



Data-driven strategies for next-generation scientific discovery in clean energy research

Citation

Häse, Florian. 2020. Data-driven strategies for next-generation scientific discovery in clean energy research. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37366005>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

Data-driven strategies for next-generation scientific discovery in clean energy research

A DISSERTATION PRESENTED

BY

FLORIAN HÄSE

TO

THE DEPARTMENT OF DEPARTMENT OF CHEMISTRY AND CHEMICAL BIOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

CHEMICAL PHYSICS

HARVARD UNIVERSITY

CAMBRIDGE, MASSACHUSETTS

MAY 2020

© 2020 – FLORIAN HÄSE

ALL RIGHTS RESERVED.

Data-driven strategies for next-generation scientific discovery in clean energy research

ABSTRACT

The transition to a low-carbon economy requires fundamental advances in clean energy technologies. Functional materials are critical to designing more efficient devices for energy generation, storage, and transmission. However, discovering energy materials is an inherently time-consuming and resource-intensive process. Current strategies are primarily based on the synergistic interplay of physical models describing macroscopic materials characteristics and experimental techniques to probe their properties. This dissertation targets the amplification of cutting edge technologies for materials discovery with data-driven tools founded on machine learning (ML). We focus on formulating restructured discovery workflows and discussing their potential to approach clean energy challenges, including natural photosynthesis and artificial light-harvesting.

In the first part, we outline the benefits of data-driven methods to the acquisition and interpretation of empirical evidence in the absence of tractable computational or experimental approaches. We demonstrate how ML tools can inexpensively estimate excitation energy transfer properties in natural light-harvesting complexes at higher accuracies than approximative physical models. We also highlight opportunities to simplify experimentation for organic electronics by identifying and exploiting statistical relations between experimentally inaccessible electronic properties from easily accessible optical properties. Finally, we showcase how ML can evidence mechanistic insights into the dynamics of light-matter interactions. We further focus on the development of data-driven tools enabling autonomous workflows to discover advanced solar cell materials. This next-generation approach to scientific discovery integrates automated platforms with data-driven methods to iteratively design, synthesize, and characterize materials candidates in closed-loops. In the second part, we formulate data-driven strategies to suggest promising materials candidates with real-time experimental feedback. In the third part, we detail algorithmic frameworks for the orchestration of autonomous workflows and demonstrate their benefits to clean energy research on two applications: (i) the discovery of conductive thin-film materials for perovskite solar cells, and (ii) the discovery of photostable polymer blends for organic photovoltaics. In both applications, the autonomous approach identified

Dissertation Advisor
Professor Alán Aspuru-Guzik

Author
Florian Häse

promising materials with more complex compositions in larger design spaces at reduced operations costs. Our findings suggest that transitioning to autonomous experimentation with predictive, intuitive, and interpretable data-driven tools can extend the boundaries of forefront technologies for the discovery of clean energy materials.

Contents

1	STRATEGIES FOR SCIENTIFIC RESEARCH	3
1.1	Introduction to the scientific method	6
1.2	Empirical trials in the scientific method	9
1.3	Data-driven scientific investigation	11
1.4	Outline of the main chapters	13
2	BACKGROUND	21
2.1	Light-harvesting as an approach to a carbon free economy	22
2.1.1	Natural light-harvesting	23
2.1.2	Theoretical descriptions of light-harvesting processes	24
2.1.3	Selected solar cell models	26
2.2	Scientific discovery with Bayesian statistics	28
2.2.1	Brief introduction to probability axioms	29
2.2.2	Derivation of Bayes' theorem	30
2.2.3	Bayesian data analysis	34
2.3	Experiment planning in the Bayesian context	43
2.3.1	Bayesian optimization for single objectives	46
2.3.2	Optimizing multiple objectives at once	49
2.4	Laboratory automation	51
2.4.1	Historical overview of laboratory automation	52
2.4.2	Advancing science with automated platforms	54
I	Machine learning in the sciences	56
3	MACHINE LEARNING FOR QUANTUM DYNAMICS	58
3.1	Traditional approaches to exciton dynamics calculations	59
3.2	Modeling of excitation energy transfer	62
3.3	Data-driven approach to excitation energy transfer	64
3.3.1	Generating the excitation energy transfer database	66
3.3.2	Principal component analysis for improved training data selection	68
3.3.3	Setup of the multi-layer perceptron architecture	68
3.4	Prediction of transfer times with neural networks	70
3.4.1	Prediction accuracies of trained multi-layer perceptrons	71
3.4.2	Comparing data-driven predictions to secular Redfield results	73
3.5	Conclusion	74
4	STUDYING CHARGE TRANSFER WITH ABSORPTION SPECTRA	76
4.1	Indirect property characterizations for organic electronics	77
4.2	Computational approaches to charge transfer studies	80
4.3	Results and discussion	87
4.3.1	Correlations between electronic spectra and coupling strengths	87
4.3.2	Machine learning models as relative classifiers	88
4.3.3	Robustness of the machine learning models	90
4.3.4	Associating machine learning models with experimental studies	91
4.4	Conclusion	92

5	MACHINE LEARNING FOR SCIENTIFIC INSIGHTS	94
5.1	Thermally activated chemiluminescence	95
5.2	Data-driven chemiluminescence	98
5.2.1	Ab initio molecular dynamics simulations	99
5.2.2	Machine learning predictions	100
5.3	Deriving mechanistic insights from data-driven findings	102
5.3.1	Validation of the dissociation time predictions	103
5.3.2	Analysis of the trained models to find correlations	104
5.3.3	Use of the trained models to test hypotheses	106
5.3.4	Ab initio molecular dynamics simulations of vibrationally excited states	108
5.4	Discussion	110
5.4.1	Interpretation of the findings of the trained models	110
5.4.2	Implications for chemiexcitation and chemical design	111
5.5	Conclusion	112

II Algorithms for closed-loop experimentation 114

6	PHOENICS: A BAYESIAN OPTIMIZER FOR CHEMISTRY	116
6.1	Data-driven strategies to experiment planning	117
6.2	Formulating Phoenix	120
6.2.1	Approximating the objective function	120
6.2.2	Acquisition function	122
6.2.3	Convergence of the approximation to the objective function	124
6.3	Results and discussion	126
6.3.1	Analytic benchmarks	127
6.3.2	Developing a collective sampling strategy	128
6.3.3	Increasing the number of dimensions	132
6.4	Applications to chemistry	133
6.5	Conclusion and Outlook	137
7	GRYFFIN: OPTIMIZATION WITH INTUITION AND INSIGHTS	139
7.1	Categorical design choices in materials science and chemistry	140
7.2	Background and related work	142
7.3	Formulating GRYFFIN	145
7.3.1	Categorical optimization with naïve GRYFFIN	146
7.3.2	Descriptor-guided searches with static GRYFFIN	147
7.3.3	Descriptor refinement with dynamic GRYFFIN	150
7.4	Synthetic benchmarks	151
7.4.1	Optimization performance	152
7.4.2	Scaling to more options and higher dimensions	153
7.4.3	Data-driven refinement of descriptors	154
7.5	Applicability of GRYFFIN to chemistry and materials science	155
7.5.1	Discovery of non-fullerene acceptor candidates for organic photovoltaics	156
7.5.2	Discovery of hybrid organic-inorganic perovskites for light-harvesting	159
7.5.3	Suzuki-Miyaura cross-coupling optimization	161
7.6	Conclusion	164

8	CHIMERA: AUTONOMOUS MULTI-OBJECTIVE OPTIMIZATION	167
8.1	Multi-objective optimization for scientific discovery	168
8.2	Formulating CHIMERA	171
8.2.1	Constructing CHIMERA	171
8.3	Performance tests on synthetic benchmarks	175
8.3.1	Deviations of the expected optimum from the actual optimum	175
8.3.2	Performance with various optimization algorithms	177
8.3.3	Behavior of optimization procedures	178
8.4	Applications of Chimera	179
8.4.1	Auto-calibrating an automated experimentation platform	180
8.4.2	Inverse-design of excitonic systems	183
8.5	Conclusions	190
III Autonomous experimentation		192
9	NEXT-GENERATION EXPERIMENTATION	194
9.1	Established and emerging approaches to experimentation	195
9.1.1	Conventional high-throughput experimentation	195
9.1.2	Navigating the candidate space with self-optimizing reactors	196
9.1.3	Self-driving laboratories	197
9.2	Challenges to the deployment of self-driving laboratories	200
9.3	Future advances of self-driving laboratories	202
9.4	The role of the researcher	205
9.5	Concluding remarks and future perspectives	205
10	ORCHESTRATING AUTONOMOUS EXPERIMENTATION	207
10.1	Autonomous approaches to scientific discovery	208
10.2	Overview of automation software in chemistry	211
10.3	Architecture of CHEMOS	214
10.4	Materials and Methods	217
10.4.1	Color, pH, density and drink experiments	217
10.4.2	Autocalibration experiment	218
10.5	Applications of CHEMOS	219
10.5.1	Orchestration of standard laboratory equipment	219
10.5.2	Autonomous calibration of a remote robotic sampling sequence	223
10.6	Conclusion	226
11	AUTONOMOUS DISCOVERY OF THIN-FILMS	228
11.1	Introduction	229
11.2	Results	231
11.3	Discussion	233
12	AUTONOMOUS DISCOVERY OF ORGANIC PHOTOVOLTAICS	235
12.1	Introduction	236
12.2	High-throughput experimentation	238
12.3	Autonomous experimentation	241
12.4	Statistical comparison and discussion	243
12.4.1	Virtual robot	244
12.5	Conclusions and Outlook	246

13 SUMMARY AND OUTLOOK	250
13.1 Conclusions	251
13.2 Future directions	255
 BIBLIOGRAPHY	259
 GLOSSARY	307

To my family. For their unmatched support.

*It is better to solve the right problem approximately
than to solve the wrong problem exactly*
– J. W. Tukey

*Auf einem Dampfer, der in die falsche Richtung fährt,
kann man nicht sehr weit in die richtige Richtung gehen.*
– M. Ende

Acknowledgments

With all the time that I enjoyed being a doctoral student on the way to pursue a Ph.D., I am convinced that this dissertation is more than just the sum of its chapters. Throughout this journey, and for everything written on the pages to follow, I have had the privilege to enjoy the advice, support, guidance, empathy, and so much more of the people surrounding me. The matter lab has been a wonderful place to grow as a scientist and as a person, and it is my distinct pleasure to share my gratitude with everyone who helped me on this path.

When I first started graduate school, I could not have imagined the countless exciting moments, adventures, and unanticipated opportunities that would emerge during the following four years. Alán Aspuru-Guzik has been a fantastic mentor at any moment of this journey and never got tired of sharing his infectious energy, excitement, and creativity. I want to thank Alán for his seemingly infinite enthusiasm and optimism, for the opportunity to explore my scientific interests, for nurturing my curiosity, for the rare openness to the ideas of a first-year graduate student, and for creating this incredibly vibrant and inspiring atmosphere in his research group. The time as a graduate student might be a time of many ups and downs, but Alán's constant support, encouragement, scientific input, strategic advice, and care for the people in his group have proven to be an invaluable asset to weather any storm. More than anything, I want to thank Alán for believing in me, both as a researcher and as a person. I am also grateful to Kang-Kuen Ni and Christopher Rycroft, who have generously offered their advice as my committee members time and again, who have consistently been an immense source of inspiration, and who provided invaluable feedback throughout my graduate studies. Kang-Kuen has always shared great perspectives on my research, helped me to focus my scientific adventures, and has supported me not only in my research but also in my role as a teaching assistant. Christopher has always helped me to advance my scientific perspectives when discussing research ideas. I deeply admire his intuitive approach to mathematical modeling and numerical analyses of physical systems. Throughout my graduate studies, I felt privileged to be mentored by a committee with such broad and profound scientific expertise and diverse insights. Thank you for all the great advice, constructive comments, and your constant support.

Throughout the years, I have witnessed the matter lab grow and shrink in size, move institutions, and explore new research directions. What has not changed, however, is the open, energetic, and creative atmosphere created by every single member of this group. From my very first days on, I

found myself fortunate to be among incredible colleagues and friends who are always open to spare some of their time and brainstorm about exciting research ideas, have the patience and kindness to share some of their wisdom, always have an open ear to listen to the inevitable challenges a graduate student learns to master over time, and never get tired to give invaluable advice. I have seen many people join and leave the group, and I am happy to say that I have enjoyed any and all interactions with everyone.

I thank Stéphanie Valteau for taking me under her wings during my first months in the group, and for patiently introducing me to the fascinating world of excitonics while tirelessly keeping me on track with the projects on which we worked. Over the years, I continuously grew my appreciation for how forgiving Stéphanie had been with all the many things first-year graduate students still need to learn. I am exceptionally grateful for her mentorship. During my time in the matter lab, Semion Saikin is probably the person with whom I shared the same office for the longest time. I want to thank Semion for the many conversations we had, for his critical and always helpful comments on my research, especially on the projects we worked on, but even more so for taking the time to also share his invaluable insights on other parts of my work. I have also had the pleasure to benefit from Christoph Kreisbeck's expertise during the early stages of my doctoral studies. I want to thank Christoph for teaching me many secrets about excitonics and computational science during our many runs up and down the Charles River. I am particularly grateful for all time that I spent with Loïc Roch on all kinds of occasions. Over the years, we have been comrades on many exciting research projects. I want to thank Loïc for innumerable discussions and coffee chats, for the runs to the Fells in the Boston summer, for the night-long writing sessions in the office, for our robotics efforts in Siberia and the hikes in the Swiss Alps. Without a doubt, Loïc had a critical impact on most of the projects presented in this dissertation. To this date, Loïc manages to push the boundaries of my imagination in our scientific efforts, and I want to thank him for his friendship. My gratitude also goes to Daniel Tabor for many great chats and discussions and introducing me to the rules and strategies of pool. Most of all, however, I want to thank Daniel for patiently listening to some ambitious research ideas and teaching me to find the balance required to transform visions into projects. During the second half of my graduate studies, I enjoyed the pleasure of engaging in countless debates with Mario Krenn, which frequently turned into fundamental discussions about the principles of science itself. I want to thank Mario for sharing his gift with me – the gift to ask the most challenging questions in the most polite way to easily spot the flaws in any idea. Mario helped me grow in my thinking and showed me the importance not just to consider the next step, but also the step after that. I

am further grateful to Matteo Aldeghi for all his support. It is a particular pleasure to brainstorm ideas with Matteo over some espresso, during which I quickly learned to appreciate the elegance with which Matteo easily simplifies seemingly unsolvable challenges into much more manageable tasks. I also want to thank Pascal Friederich and Gabriel dos Passos Gomes for many exciting discussions about chemistry, photovoltaics, life, and anything in between. I appreciate all the kindness they shared with me at any moment.

My fantastic experience in the matter lab has much to do with the great fellow graduate students I have the privilege to call friends. Many thanks to Hannah Sim, Tim Menke, Jennifer Wei, Teresa Tamayo-Mendoza, Riley Hickman, Jhonathan Romero, Nicolas Sawaya, Benjamín Sanchez-Lengeling and Adrián Jinich for all the fun moments we shared, for all the interactions that I believe only happen between graduate students, for your constant support, the laughter, the great science we did together and the times we spent outside the lab. At some point or another, I have also immensely benefitted from interactions with Doran Bennett, Matthias Degroote, Si Yue Guo, Martha Flores, Abhinav Anand, Akshat Nigam, Cyrille Lavine, José Darío Perea and everyone else in the matter lab. I would further like to extend my gratitude to the amazing laboratory administrators who never got tired to answer my many questions, always found creative solutions for any unusual request, never complained when I checked in on the status of a pending reimbursement, and always helped me to find a slot in Alán's schedule: Cynthia Chew, Marlon Cummings, Siria Serrano, Felixander Negrón, Mukesh Dodain, and Irene Zuniga.

During my time in the matter lab, I enjoyed plenty of runs and races in New England and Ontario with many of the Quantum Tunnelers. Our frequent runs up and down the Charles River, the Middlesex Fells, Mount Monadnock, Yellow and Mud Creek, or Tommy Thompson Park have been great distractions to escape the office and breath some fresh air.

For my wonderful experience at the Department of Chemistry and Chemical Biology, I would like to thank the Department staff for their support. I am grateful for all the help I received, especially from Joe Lavin and Kathy Oakley, who helped me navigate the administrative hurdles of a graduate program. I want to thank the many students in PS50 and Chem160 who embarked with me on exciting journeys to discover the principles of scientific computing and quantum mechanics. I would also like to acknowledge the tireless support of the Research Computing Group of the FAS Division of Science at Harvard, and the StackExchange and StackOverflow communities. They never left a single question unanswered. I want to further extend my gratitude to the Herchel Smith Fund and the Jacques-Emile Dubois Fund for their financial support during my graduate studies and

the many stimulating interdisciplinary symposia. In this spirit, I would also like to acknowledge the German Academic Scholarship Foundation, whose financial support allowed me to focus on my undergraduate studies. This dissertation would not have been possible without them.

I also want to thank my collaborators in the groups of Prof. Christoph Brabec, Prof. Roland Lindh, Prof. Jason Hein, Prof. Curtis Berlinguette, and Prof. Christopher Cheatum for the many exciting research questions that we approached together.

Walking along the path of pursuing a Ph.D. is the result of the encouragement and support I received from inspirational teachers in the past. I would like to specifically mention my undergraduate advisors, Martin Zacharias and Matthias Rief, for their encouragement and the opportunities they offered. I am also grateful to Carsten Rohr for sharing his advice and his wisdom about higher education. Finally, I would like to thank my high school math teacher Dörthe Moll, whose rigorous analytic thinking about problems in mathematics and beyond influences me in my research to this day.

In this journey, I also had the pleasure to enjoy the support of dear friends who have been a constant source of encouragement and helped me keep my feet on the ground. Among others, I want to specifically thank Maria, Johanna, Benjamin, and Miguel for the many fun moments we shared. Most importantly, I want to thank my parents, brother, and grandparents. Thank you for fostering my interest in science, for never getting tired to respond to the questions one can only ask at a young age and for encouraging me to explore my interests. Your everlasting support throughout every step of my life is invaluable. Finally, I thank you, Sophie, for all the joy and laughter, for your patience, for the time spent together in Cambridge, New Haven, Frankfurt, Toronto, Berlin and anywhere in between, and for always having been present regardless of the distance.

Citations to previously published work

At the moment of submitting this dissertation, the contents of chapters 3 to 12 have, apart from minor modifications, appeared as published articles or preprints. I acknowledge the scientific contributions of all the authors of these publications to the present document.

Florian Häse, Christoph Kreisbeck and Alán Aspuru-Guzik. Machine learning for quantum dynamics: deep learning of excitation energy transfer properties. *Chem. Sci.* **8** (12), 8419–8426 (2017).

Loïc M. Roch, Semion K. Saikin, Florian Häse, Pascal Friederich, Randall H. Goldsmith, Salvador León and Alán Aspuru-Guzik. From absorption spectra to charge transfer in nanoaggregates of oligomers with machine learning. *ACS Nano*. in press, 10.1021/acsnano.0c00384 (2020).

Florian Häse, Ignacio Fdez. Galván, Alán Aspuru-Guzik, Roland Lindh and Morgane Vacher. How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry. *Chem. Sci.* **10** (8), 2298–2307 (2019).

Florian Häse, Loïc M. Roch, Christoph Kreisbeck and Alán Aspuru-Guzik. Phoenix: A Bayesian optimizer for chemistry. *ACS Cent. Sci.* **4** (9), 1134–1145 (2018).

Florian Häse, Loïc M. Roch and Alán Aspuru-Guzik. Gryffin: An algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications to chemistry. *arXiv preprint*. arXiv:2003.12127 (2020).

Florian Häse, Loïc M. Roch and Alán Aspuru-Guzik. Chimera: enabling hierarchy base multi-objective optimization for self-driving laboratories. *Chem. Sci.* **9** (39), 7642–7655 (2018).

Florian Häse, Loïc M. Roch and Alán Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. *Trends Chem.* **1** (3), 282–291 (2019).

Loïc M. Roch*, Florian Häse*, Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars P. E. Yunker, Jason E. Hein and Alán Aspuru-Guzik. ChemOS: orchestrating autonomous experimentation. *Sci. Robot.* **3** (9), eaaat5559 (2018).

Loïc M. Roch*, Florian Häse*, Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars P. E. Yunker, Jason E. Hein and Alán Aspuru-Guzik. ChemOS: An orchestration software to democratize autonomous discovery. *PLoS one* **15** (4), e0229862 (2020).

Benjamin P. MacLeod, Fraser G. L. Parlane, Thomas D. Morrissey, Florian Häse, Loïc M. Roch, Kevan E. Dettelbach, Raphaell Moreira, Lars P. E. Yunker, Michael B. Rooney, Joseph R. Deeth, Veronica Lai, Gordon J. Ng, Henry Situ, Ray H. Zhang, Alán Aspuru-Guzik, Jason E. Hein, Curtis P. Berlinguette. Self-driving laboratory for accelerated discovery of thin-films. *arXiv preprint* arXiv:1906.05398 (2019).

Stefan Langner*, Florian Häse*, José Darío Perea, Tobias Stubhan, Jens Hauch, Loïc M. Roch, Thomas Heumueller, Alán Aspuru-Guzik and Christoph Brabec. Beyond ternary OPV: High-throughput experimentation and self-driving laboratories optimize multi-component systems *Adv. Mater.* **32**, 1907801 (2019).

* These authors contributed equally

Throughout the course of my doctorate, I also contributed to the papers and preprints listed below, which are not explicitly included as independent chapters.

Florian Häse, Teresa Tamayo-Mendoza, Carmen Boixo, Jonathan Romero, Loïc M. Roch, Alán Aspuru-Guzik. Autonomous titration for chemistry classrooms: preparing students for digitized chemistry laboratories. *chemRxiv preprint*. 10.26434/chemrxiv.12097908.v1 (2020).

Tim Menke, Florian Häse, Simon Gustavsson, Andrew J. Kerman, William D. Olliver, Alán Aspuru-Guzik. Automated discovery of superconducting circuits and its application to 4-local coupler design. *arXiv preprint* arXiv:1912.03322 (2019).

Philip Pagano, Qi Guo, Chetya Ranasinghe, Evan Schroeder, Kevin Robben, Florian Häse, Hepeng Ye, Kyle Wickersham, Alán Aspuru-Guzik, Dan Major, Lokesh Gakhar, Amnon Kohen and Christopher Cheatum. Oscillatory Active-Site Motions Correlate with Kinetic Isotope Effects in Formate Dehydrogenase. *ACS Catal.* **9** (12), 11199–11206 (2019).

Santiago Vargas, Siavash Zamirpour, Shreya Menon, Arielle Rothman, Florian Häse, Teresa Tamayo-Mendoza, Jonathan Romero, Sukin Sim, Tim Menke, Alán Aspuru-Guzik. Team-based learning for scientific computing and automated experimentation: Visualization of colored reactions. *J. Chem. Ed.* **97** (3), 689–694 (2020).

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich and Alán Aspuru-Guzik. SELFIES: A robust representation of semantically constrained graphs with an example application in chemistry. *arXiv preprint* arXiv:1905.13741. 2019.

Stéphanie Valleau, Romain A. Studer, Florian Häse, Christoph Kreisbeck, Rafael G. Saer, Robert E. Blankenship, Eugene I. Shakhnovich and Alán Aspuru-Guzik. Absence of Selection for Quantum Coherence in the Fenna-Matthews-Olson Complex: A combined evolutionary and excitonic study. *ACS Cent. Sci.* **3** (10), 1086–1095. 2017.

The following software packages and data repositories have been released as part of this dissertation:

GRYFFIN: An algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications to chemistry (2020).
<https://github.com/aspuru-guzik-group/gryffin>.

SCILLA: Automated discovery of superconducting circuits and its application to 4-local coupler design (2019).
<https://github.com/aspuru-guzik-group/scilla>.

CHIMERA: Enabling hierarchy-based multi-objective optimization for self-driving laboratories (2018).
<https://github.com/aspuru-guzik-group/phoenics>.

PHOENICS: A Bayesian optimizer for chemistry (2018).
<https://github.com/aspuru-guzik-group/phoenics>.

CHEMOS: Orchestrating self-driving laboratories (2018).
<https://github.com/aspuru-guzik-group/ChemOS>.

Deep learning of excitation energy transfer properties at Redfield accuracy (2017).
<https://github.com/FlorianHase/LearningExcitonTransfer>.

Introduction

This page is intentionally left blank.

1

Strategies for scientific research

Humanity strives to explore the world, to understand its fundamental governing principles, and to evaluate how the laws of nature can be leveraged to advance society. Science has always been at the heart of these efforts. In our modern understanding, scientific research entails systematic investigations of causal relations and the acquisition of knowledge about nature and society. To this end, scientists aspire to conceive hypotheses, collect observations, organize knowledge and formulate insights about the systems they study, intending to conceptualize inexplicable phenomena and devise testable predictions about the world. Scientific discoveries have brought immense benefits and shape our daily lives in several facets.

Fundamental advances in biology, physics, and chemistry can stimulate such breakthroughs in medicine, technology. The discovery of insulin, for example, as a robust and reliable treatment for diabetes mellitus, was enabled by foundational studies on the role of pancreatic extracts on blood sugar levels in the early 20th century.^{1,2} Based on these pioneering findings, Frederick Banting in collaboration with Charles Best and John Macleod demonstrated the isolation and controlled production of insulin in 1922.³ The first injections of the newly discovered insulin extract confirmed its use to treat diabetes mellitus,⁴ and quickly prompted its industrial production by the pharmaceutical company Eli Lilly. Several human and animal-source analogs of insulin are available today and are considered to be among the safest and most effective medications regularly used by millions of people. The discovery of the transistor by John Bardeen, Walter Brattain and William Shockley in 1947⁵ marked a scientific advance leading to remarkable technological developments. Transistors are semiconductor devices which enable the control of an electrical signal *via* an external voltage. The functionality to amplify and switch electrical signals had already been demonstrated with vacuum tubes in the early 20th century. However, the point-contact transistor developed by Bardeen, Brattain, and Shockley presented revolutionary improvements over vacuum tube designs regarding size, power consumption, and overall reliability and durability. Thus, the transistor suddenly provided a

substantially more viable design to the development of integrated circuits for microprocessors and eventually paved the way to modern classical computing. But single scientific achievements can also benefit multiple disciplines at once. For instance, the successful reconstruction of the light-harvesting reaction center *via* X-ray crystallography by Johann Deisenhofer, Hartmut Michel, and Robert Huber in 1984⁶ offered groundbreaking insights into natural photosynthesis with new technological inventions to study membrane proteins. Resolving the reaction center at the atomic level catalyzed our microscopic understanding of natural solar energy conversion. The identified principles of photosynthetic processes in plants and bacteria sparked the development of artificial clean energy technologies, particularly nature-inspired light-harvesting devices. But the work of Deisenhofer, Michel, and Huber also benefitted further investigations on other membrane proteins, as their detergent-based protocol for the crystallization of the reaction center proved to become a critical and widely transferable tool.

The revolutionary studies mentioned in the previous paragraph arguably provided immeasurable benefits to society. Yet, this century still poses critical open challenges, ranging from climate change over food security to global health. The development of solutions to these challenges involves the discovery of novel materials for renewable energy, environmentally friendly pesticides and next-generation antibiotics and requires advanced approaches to scientific investigation. These efforts can be categorized into basic and applied science based on their intentions and scopes. Basic science focuses on the investigation of the fundamental outstanding questions about nature, motivated by intrinsic curiosity and the desire to complete our understanding its governing principles. To that end, basic science is mostly hypotheses-driven and targets the formulation and validation of physical models. Applied science, in contrast, is pursued to leverage existing scientific knowledge to approach practical problems by devising technologies built upon established principles to control and modulate the properties of physical systems. Both of these intentions are highly complementary with basic research providing the critical foundations to fuel innovations of practical relevance. The discovery of advanced materials, for example, is recognized as one of the critical efforts which requires substantial advances in both basic and applied science. Advanced materials already affect some aspects of society today,⁷ including clean energy generation with solar cell technologies, energy transmission and storage *via* battery solutions, or water filtration with novel catalysts.^{8–16} Yet, the discovery of such materials is heavily involved and to date requires significant amounts of resources over several years of research before materials technologies reach the market. One of the fundamental challenges in materials discovery, particularly for clean energy technologies, consists in

understanding the causal relations based on which macroscopic materials properties emerge from microscopic phenomena. Complete and comprehensive theoretical approaches to model materials properties are rarely computationally tractable due to the many degrees of freedom in these systems, while dominant microscopic features might be unobservable with experimental approaches. In this dissertation, we explore opportunities to amplify cutting edge theoretical models and experimental technologies with data-driven approaches to enable more elaborate and more streamlined workflows for scientific discovery with a focus on clean energy technologies to face some of the immediate societal challenges.

Scientific investigation follows a discovery process which primarily relies on the interplay of systematic observation and experimentation, inductive and deductive reasoning, and the formation of testable hypotheses and theories.¹⁷ In some cases, this iterative process can be initiated by a fortuitous event, such as the discovery of X-radiation and its implications for medical diagnostics. In 1895, Wilhelm Conrad Röntgen was working with an X-ray emitting cathode ray tube, when he realized that a nearby fluorescent screen would glow when the tube was on, even though it was covered.¹⁸ While trying to block the tube further, Röntgen realized that an image of the bones his hand was projected on the screen whenever he placed his hand directly between the tube and the screen. A similarly coincidental scientific advancement was the discovery of penicillin. This group of antibiotics derived from common molds was first observed and described by Alexander Fleming in 1928. Although Fleming had already been studying the antibacterial properties of different types of molds, it was an accident that focused his attention specifically on penicillin. A Petri dish with *Staphylococci* had been mistakenly left open. Fleming found that it got contaminated by mold, which he later correctly classified as penicillin. He described a halo of inhibited bacteria growth around the mold, which motivated him to analyze it in more detail and eventually describe penicillin as a potent antibiotic.¹⁹ Although the initial observations in these two examples might have been incidental and without the intent to provoke discoveries of such impact, a substantial amount of further reasoning and experimenting was needed to quantify and conceptualize these findings. In fact, most of today's technologies, such as modern computing and treatments for fatal diseases, have primarily emerged from our flourishing understanding of the physical and the life sciences. This interplay between systematic observation and theoretical reasoning is commonly summarized in the scientific method.

1.1 INTRODUCTION TO THE SCIENTIFIC METHOD

Advancements in science are driven by the cooperative interaction of two fundamental approaches to gaining insights: (i) empirical trials to systematically probe and observe the governing principles of nature, and (ii) theoretical modeling and hypothesizing to describe physical systems, conceptualize empirical findings, and to make testable predictions. Scientists across all fields routinely conceive and revise theoretical models based on empirical evidence in an iterative, cyclical process referred to as the *scientific method*.^{20,21} The scientific method includes a series of tools, techniques, and best practices for systematic investigation. It further spans the formulation of hypotheses *via* induction based on the observations and the thorough testing of these hypotheses with possibly additional experimental evidence and measurements. Essential to this process is that researchers remain unbiased in their interpretations of collected observations by applying rigorous skepticism to derive the most plausible conclusions supported by the empirical evidence. Although different formulations of the scientific method exist, it is generally recognized to involve empirical observations about a studied system and is comprised of variations of four indispensable steps, which can be followed in an arbitrary order:^{22,23}

- (i) **EXPERIMENTING:** The experimentation step targets the collection of empirical evidence to unveil governing principles. To that end, the studied system is systematically observed by measuring its responses upon controlled interactions with the environment.
- (ii) **CHARACTERIZING:** Given empirical evidence of the properties or the behavior of a studied system, this step aims to provide the context for these observations. Accordingly, the characterization can require profound statistical analyses of the collected evidence and the uncertainties about fluctuations in the measurements.
- (iii) **HYPOTHESIZING:** Inspired by the characterized observations, this step focuses on the development of physical models to describe causal relations determining the behavior of the studied system. The hypothesizing step could evaluate the level of agreement between predictions of existing models (see next step) and the collected evidence. However, the hypothesizing step can also be the start of the investigation cycle when conceiving theoretical models to describe physical systems before collecting empirical evidence.
- (iv) **PREDICTING:** Following the formulation of a physical model to explain the empirical evidence, the prediction step uses this model to make testable predictions for future experimentation *via*

inductive reasoning.

Although the term *scientific method* only emerged in the 19th century, when science was increasingly institutionalized with a focus on the introduction of defining terminologies,²⁴ this inquiry strategy to gain scientific understanding has been applied throughout the centuries. The process typically requires multiple iterations of generating hypotheses and conducting experiments before satisfyingly accurate theories are identified. In this process, it is not uncommon that individual steps are carried out by different researchers. In some cases, the scientific method is followed implicitly over decades, questioning established pictures of nature and re-evaluating prior assumptions. For instance, more than 2,000 years ago, Aristotle introduced an early theory of gravity based on observations he made at his time, which states that massive objects fall at a speed that is proportional to their masses. This hypothesis has been questioned on multiple occasions by several scholars, including Galileo Galilei, who conducted a series of experiments in 1591. Galilei is said to have dropped geometrically identical spheres of different masses from the top of the leaning tower of Pisa.^{25,26} His measurements indicated that there is no significant difference in the times at which the spheres hit the ground, based on which he concluded that Aristotle's theory of gravity is incorrect. Instead, he suggested a *law of odd numbers*, which states that the distances which the spheres travel are proportional to the squares of the elapsed times. While this phenomenological law seemed to provide reliable predictions about the fall of massive objects, it was not before 1687, when Isaac Newton's introduced his Second Law and Law of Gravity, that a physical model provided a causal explanation and quantitatively accurate predictions for the fall of massive bodies.²⁷

The iterative character and inherent skepticism are essential components of the scientific method and fundamental to judging how well physical models describe experimental evidence. This critical point is illustrated by the Kaufmann-Bucherer-Neumann experiments, conducted at the beginning of the 20th century, which aimed to measure the dependence of the inertial mass of an electron on its velocity. Towards the end of the 19th century, it had been speculated that the mass of the electron increases with its velocity. Walter Kaufmann started a series of experiments in 1901 to test this hypothesis. Kaufmann used radium as a source of beta radiation to generate electron beams where electrons could reach speeds of up to $0.9 c$. He homogenized the speed of the electrons in the beam with a Wien filter,²⁸ and subsequently bent the path of the electrons with carefully arranged electromagnetic fields.²⁹ The curvature of the bent electron beam is modulated by the charge to mass ratio of the electrons, which indirectly enabled measurements of the dependence of the elec-

tron's mass on its velocity. Kaufmann's experiments confirmed the theoretical speculation that the charge to mass ratio indeed decreases with increasing speed. At the time that Kaufmann published his results, multiple competing physical models to describe such a decrease had been introduced, including models from George Searle,³⁰ and Hendrik Antoon Lorentz.³¹ Kaufmann concluded that his experimental observations did not agree with any of these models. This disagreement sparked the construction of further physical models, such as the one introduced by Max Abraham in 1902.³² Abraham assumed that the electron is a rigid sphere of finite volume where the charge is distributed evenly across the surface. He expressed the electronic mass as a combination of longitudinal and transversal masses with respect to the direction of movement of the electron. Lorentz also refined his theory by the assumption that the charge of the electron was distributed throughout the sphere and that the sphere would be compressed longitudinally.³³ Further refinements of this theory by Poincare would yield results which were quantitatively in agreement with the later established principle of relativity.³⁴ Kaufmann also refined his experimental apparatus to reduce uncertainties in the observations. He published an update to his experiment in 1902, where he corrected mistakes in his calculations and repeated the experiments at higher statistical confidence.³⁵ Despite his updates, the apparatus of Kaufmann still had some significant measurement uncertainties. With these uncertainties, Kaufmann's analysis suggested that Abraham's theory would be correct, thus rejecting the mathematically correct expression of Lorentz. But Max Planck repeated Kaufmann's analysis in 1906 to find that the analysis was inconclusive, such that neither of the theories would agree to a satisfactory degree with the experimental observations.³⁶ Meanwhile, in 1905, Albert Einstein introduced his theory of special relativity, which he used to explain the mass increase of the electron based on transformations between the rest frame of the particle and the laboratory frame. In his theory, the electron was described as a point charge, thus questioning the geometry of the electron even more profoundly than previous theories.³⁷ The mass increase of the electron derived from special relativity is mathematically equivalent to the expression derived by Lorentz, despite substantial differences in the underlying physics, such that Einstein's theory did not agree with Kaufmann's results. Alfred Bucherer went on to improve Kaufmann's experimental apparatus in 1909 by replacing the parallel magnetic and electric fields with perpendicular magnetic and electric fields to bend the electron beam.³⁸ Calculating the charge to mass ratio at rest from electrons at different speeds, Bucherer found that only the Lorentz-Einstein equation gave consistent results and concluded that his experimental evidence favors this theory over Abraham's model. The apparatus was further refined by Günther Neumann in 1914, whose results again confirmed Lorentz-Einstein.³⁹ Further ex-

periments and analyses aimed to distinguish between the competing theories, and multiple disputes were held to discuss experimental setups and measurement analyses, but special relativity eventually emerged as the most predictive physical model, also because of accurate predictions for other problems in physics at that time. These discussions, however, demonstrate that systematic observation and experimentation, inductive and deductive reasoning, the formation of testable hypotheses and theories, and a certain amount of skepticism towards observations and interpretations are at the heart of scientific research and the essential tools to drive scientific understanding.

1.2 EMPIRICAL TRIALS IN THE SCIENTIFIC METHOD

Theoretical hypotheses can drive our quantitative understanding of nature's governing principles and the development of physical models to predict the state or behavior of a studied system. However, the collection of empirical evidence from experiments is critical to the scientific method as it provides insights into prevalent interactions between physical systems. Experiments in the natural sciences can be conducted in different settings with various tools. The most conventional form of an experiment might be the laboratory experiment, where physical systems are prepared in a well-defined initial state, and their behavior and properties are observed over time under controlled environmental influences and interactions. In some instances, the careful manipulation of the system to prepare the initial state is not possible, for example in astronomical studies where experiments mostly rely only on observations. Computational tools can also provide insights into the credibility of a given physical model by generating empirical data within the possibilities and limitations of the computational model.⁴⁰ To that end, computational models are typically validated *via* the reproduction of observations obtained in laboratory experiments. Assuming that predictions of the computational model agree with the laboratory observations, further empirical evidence can be collected computationally, which is otherwise inaccessible to laboratory experiments. For example, atomistic simulations of biomolecular many-particle systems can reveal the governing mechanisms of vital metabolic processes at length and time scales which cannot be reached experimentally.⁴¹ Yet, computational experiments do not necessarily model the physical reality, and information extract from computational experiments is only valid with respect to the computational model from which it has been generated. Experiments in engineering can resemble experiments in the natural sciences, for example, in nondestructive testing where properties of materials or components are measured and evaluated. However, while experiments in the natural sciences tend to study causal relations be-

tween physical systems, experiments in engineering are mostly executed to confirm the final outcome of a mechanical or electrical process. Throughout this dissertation, we will use the term *experiment* to collectively refer to any of these three types of experiments, *i.e.*, laboratory experiments, computational experiments and nondestructive testing, and, more generally, to any method which probes hypotheses in the broader context.

Crucial to the empirical testing of a hypothesis is the degree of agreement between theoretical predictions and empirical observations. Experiments require precise measurement instruments and well-defined setups, *i.e.* carefully prepared and clearly described initial states of the physical system, which is observed and studied over time. The effort and resources required to prepare this initial state can vary widely between experiments. Measuring the frequency of a resonant LC circuit, for example, can be as simple as connecting an inductor and a capacitor to an electric circuit and using a voltmeter to measure the voltage over time. Other setups can be more elaborate by several orders of magnitude, such as the Large Hadron Collider (LHC) which was constructed to test different theories of particle physics and constitutes the highest-energy particle collider to date.⁴² For an experiment to represent an adequate test of a physical model, and to be considered sufficiently significant to reject a hypothesis or decide between theories, the experiment is typically repeated with different setups and by different researchers. A detailed description of the experimental setup is therefore as essential as the diligent execution of the experiment to achieve accurate and reproducible results.

Even with precisely controlled initial states and measurement procedures, experiments are subject to observational errors which can arise due to many factors. These errors are not to be confused with mistakes but are rather an inherent part of the experimental setup or the measurement process, which cannot be further reduced or entirely avoided. Understanding the accuracy and reproducibility of the experimental setup is of fundamental importance, especially when comparing theoretical models, to avoid misinterpretations as in the case of Kaufmann (see Sec. 1.1). An agreement between the predictions and the measurements still does not prove that the theory is true, but only supports the theory. The example of Lorentz (see Sec. 1.1) illustrates that a physical model can produce numerically accurate predictions for specific experiments. Nevertheless, such a theory might be replaced in favor of another physical model, which also accurately predicts the outcomes of other experiments where properties of different systems are measured. If the observation of the collected measurements is highly unlikely under the inspected theory even with sources of noise much larger than those expected to be present in the experiment, the experiment might not support the theory. Thus, conclusive statistical analyses of experimental results must always be guided by a notion

of expected error on the collected observations. Measurements can be modulated by systematic and random observational errors potentially introduced both at the time of the preparation of the initial state and at the time of the measurement. Systematic errors describe deviations in the measurements from their expected physical values which are caused by an inaccuracy inherent to the experimental setup. As such, a systematic error will reproducibly modulate the experimental outcome such that measurements subject to systematic errors consistently over- or underestimate the expected values. Random errors instead are stochastic and are expected to average out for multiple repetitions. Experiments usually present with both sources of errors, and a detailed description of the experimental setup is required to recognize all relevant factors which could modulate the experimental outcome. Omitting or not recognizing the influence of an environmental variable, such as the humidity in the laboratory, can result in seemingly unexplainable fluctuations in the measurements.^{43,44}

Empirical trials for hypothesis testing generally comprise the more resource-demanding step of the scientific method and need to balance measurement accuracies and confidence in the empirical evidence with a given budget for the required resources. Budgets can be constrained by a variety of factors concerning both the experimental setup and the measurement procedures. These constraints can include time and experimentation throughput, money to purchase laboratory equipment and consumables, or even opportunity costs and safety aspects. Planning an experimental setup requires to compensate several of these constraints, which in some cases can impede the scope or even the feasibility of an experimental study. Innovation on experimental equipment and tools to lower these obstacles is imperative to advance scientific research and enable more comprehensive investigations, as it was demonstrated for obtaining the crystal structure of the reaction center.⁶ The discovery of functional materials for clean energy technologies is one example where the discovery of viable materials solutions can require up to two decades of basic and applied research, and tools to lower the demands crucial for a timely transition to a low-carbon economy.

1.3 DATA-DRIVEN SCIENTIFIC INVESTIGATION

The scientific method relies on an iterative process of suggesting, building, testing, discarding and, refining physical models to describe the causal relations governing the processes observed in nature (see Sec. 1.1).⁴⁵ Experiments provide empirical evidence to test and validate the conceived models but can be highly resource-demanding and, in some cases, not realizable without limitations arising

from constrained budgets (see Sec. 1.2). Thoroughly analyzing and conceptualizing experimental observations constitutes a crucial step of this procedure. It is through statistical approaches that we quantify the correspondence between model predictions and experimental observations while accounting for measurement accuracies and hidden factors.⁴⁶ Statistical modeling has, therefore, always been at the heart of scientific investigations. Although the statistical analysis of collected data is often treated as the final subsidiary step to a successful experiment, the ambition to extract robust statistical statements can influence the design and strategy of the experiment.⁴⁷ Statistical tools can, therefore, be considered as an opportunity to maximize the information gain from empirical evidence and thus catalyze scientific research by lowering the requirements on experimental trials and inspire promising physical models to describe causal relations detected from empirical correlations. To that end, artificial intelligence (AI), notably machine learning (ML), experience rising interest by the scientific community^{48–60} motivated by remarkable successes on complex tasks such as image recognition,⁶¹ natural language processing,⁶² or strategic decision making.⁶³

The fields of ML and statistics share many similarities, and both statistical methods and ML models may be used for inference and prediction to make quantitative statements about physical systems. However, both fields also present subtle differences with individual advantages and limitations. The appeal to use statistical methods for scientific investigation is mostly established from their robust modeling and inference capabilities. While statistical models are often based on prior assumptions about the physical reality, their main focus is on the identification of empirical relations between physical variables and the significance of these relations. As statistical models are explicitly condition on existing knowledge and expectations, their predictions are generally amenable to further interpretations. ML models, on the other hand, leverage general-purpose non-parametric algorithms which require minimal assumptions about the data-generating system and can model highly non-linear interactions at high prediction accuracies from data alone.⁶⁴ As such, ML collectively comprises data-driven approaches to extract patterns from data based on the empirical correlations to achieve high prediction accuracies on a given task. By alleviating the requirements of prior assumptions about the studied system, techniques developed in the field of ML promise opportunities to automate the data analysis and inference steps of scientific investigation. In this process, ML models rely on fundamental concepts of statistics. The boundaries between statistical inference and ML are, therefore, sometimes blurry.

Data-driven models feature one aspect, which fundamentally sets them apart from physical models and could be understood as a strength and a weakness alike. Physical models aspire to describe

an underlying reality and can be wrong in their description. Data-driven models, however, are constructed based on empirical correlations and cannot be wrong about the empirical evidence. Albeit, they cannot be right about the physical reality either. Instead, data-driven models intend to be useful for a given task and could be highly accurate or poor and inconclusive in this process. It is essential to acknowledge this aspect when incorporating data-driven tools into scientific research and to leverage this feature as an advantage. In this dissertation, we explore opportunities to benefit from these strengths of data-driven approaches in every individual step of the scientific method to enhance and amplify clean energy research.

1.4 OUTLINE OF THE MAIN CHAPTERS

This dissertation is divided into three parts, which collectively explore opportunities to enhance scientific research on clean energy technologies with data-driven approaches. In the first part, we suggest and discuss possibilities to leverage the predictive capabilities of data-driven strategies in the absence of physical models to lower the resource demands of computational and experimental studies. In this spirit, we specifically follow the goal to facilitate the inexpensive large-scale collection of empirical evidence, to probe experimentally inaccessible properties of physical systems, and to inspire physical insights and scientific understanding of molecular mechanisms. This part specifically focuses on the augmentation of the CHARACTERIZING and HYPOTHESIZING steps of the scientific method with data-driven approaches (see Sec. 1.1). The second part targets the vision to streamline scientific discovery, especially in the context of discovering functional molecules and advanced materials for artificial light-harvesting, by introducing algorithmic approaches for data-driven experiment planning in autonomous workflows based on computations and experiments. To that end, we explore how the PREDICTING step of the scientific method can be executed with data-driven tools. The third part extends this vision by introducing the layout of control software to orchestrate autonomous experimentation workflows with minimal human intervention. This final part demonstrates developments towards the digitization of all steps of the scientific method and showcases the applicability and the benefits of autonomous experimentation on two different applications on clean energy technologies.

PART I: MACHINE LEARNING IN THE SCIENCES

ML is emerging as a versatile tool for scientific research for light-harvesting technologies on various levels,⁶⁵ notably for the rapid screening of materials properties,^{66,67} candidate selection,^{68–70} results analysis,⁷¹ and interpretation.^{72,73} With the capacity to construct inexpensive predictive models from empirical data, ML technologies find applications in areas where profound physical models are not known, computational approaches are intractable, and experimental investigations are too time-consuming or resource-demanding. In this part, we outline possibilities to augment conventional approaches to scientific research with data-driven techniques in several aspects, including straightforward property prediction tasks but also going further towards the flexible integration of proxy properties into scientific workflows, and inspiring physical insights and chemical concepts from collected data. In this process, we directly exploit the capacity of ML models to make testable predictions based on empirical evidence, intending to estimate selected properties of physical systems at a high degree of accuracy. We will demonstrate that the capacity of ML to infer patterns from data can have a substantial impact on the design of scientific studies as it enables novel routes to knowledge acquisition.

In Chapter 3, we present a purely predictive application of ML models as tools to obtain accurate, low-cost estimates of the excitation energy transfer (EET) properties of excitonic systems as they are found in natural photosynthesis and artificial light-harvesting technologies. Understanding the structure-property relation of excitonic networks is recognized to be of fundamental importance to the design of novel light-harvesting technologies (see Sec. 2.1). Yet, experimental approaches and computational models to study EET properties of excitonic systems are often resource-demanding or numerically involved, and can only be afforded for selected systems. This chapter introduces a deep learning architecture based on feedforward neural networks, which predict EET properties of excitonic systems on the millisecond timescale at accuracies comparable to approximative physical models, which require several minutes of evaluation time. The data-driven models developed in this chapter, therefore, enable the large-scale investigation of EET characteristics of excitonic systems with vastly different structures to unveil the causal structure-property relations and inspire future design choices.

Chapter 4 extends the idea of using ML models to predict the properties of physical systems for light-harvesting applications to enable large-scale data acquisition experimentally. Here, we focus on organic electronics applications relevant to solar cell architectures. Specifically, we consider

the organic polymer mixture poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS), which is one of the cornerstone materials in organic electronics due to its high conductivity. However, PEDOT:PSS presents with a highly disordered morphology on multiple length scales, which modulates its electronic properties. While the microscopic structure of molecule and polymer packing is challenging to measure directly, optical spectra can be easily obtained experimentally. We suggest that data-driven approaches can enable indirect characterizations of electronic properties *via* optical measurements, which are significantly more accessible experimentally. While physical models relating absorption spectra to intermolecular couplings are unknown, and a strict causal relation is unlikely to exist, we demonstrate that our data-driven approach can indeed identify correlations between optical and electronic properties of PEDOT:PSS blends without using additional structural information which could not be obtained in experiments. The data-driven models constructed in this chapter are based on Bayesian formulations of feedforward and convolutional neural networks (CNNs), particularly suited for robust predictions with little data. The ML models constructed in this study are capable of estimating electronic couplings between PEDOT:PSS dimers from their electronic absorption spectra. With this result, this study constitutes an example where ML technologies can be used to enable experimental investigations of the conductive properties of a material *via* measurements of other properties, based on identifying and exploiting a property-property relation.

Finally, Chapter 5 demonstrates how data-driven tools can be used to spark physical understanding and inspire mechanistic insights into the dynamical behavior of a physical system. As a model system, we study the thermally activated chemiluminescent dissociation of 1,2-dioxetane. The prevalent nuclear motions governing the dissociation mechanism in this compound are highly debated, with much speculation about molecular modifications to modulate dissociation time scales. We suggest that data-driven approaches are not only capable of implicitly identifying the relevant nuclear coordinates from a sparse set of computationally simulated dissociation trajectories, but also explicitly evidence the dominant collective motions which promote or delay the dissociation. Robust predictions at high accuracies are achieved with Bayesian feedforward neural networks. We demonstrate how the architecture of the trained ML models can be analyzed to promote physical insights and how the trained models can be deliberately used to confirm or reject hypotheses regarding the dissociation mechanism which are inaccessible experimentally and intractable computationally.

PART II: ALGORITHMS FOR CLOSED-LOOP EXPERIMENTATION

The scientific method provides a framework for scientific research founded on the interplay between the conception of physical models hypothesizing about the governing principles of nature, and the empirical investigation of physical systems to study the causal relationships between them (see Sec. 1.1). In Part I, we mostly focused on the predictive capabilities of data-driven approaches to make quantitative predictions about physical systems and inspire scientific insights during this process. Parts II and III instead explore the benefits of the data-driven modeling capacities of ML tools to scientific research, specifically to scientific discoveries in a slightly narrower context. Although scientific discovery broadly describes any successful scientific inquiries, it can also be related to questioning the existence of a physical system with desired properties, which is a task commonly encountered in the applied sciences or engineering. Prominent examples of this narrower context include drug discovery and materials discovery.

In this context, scientific discovery is closely related to inverse-design tasks, which aim to identify the structural features of a physical system from which desired properties emerge. While data-driven generative inverse-design workflows are emerging,⁷⁴⁻⁷⁷ the dominant strategy for such tasks in the absence of complete and comprehensive theoretical descriptions consists of undirected empirical searches. Such Edisonian trial-and-error approaches are based on the explicit or implicit construction of a pool of promising materials candidates. The generation of this candidate library can be inspired from theoretical expectations, experimental evidence, or constraints due to limited available resources, and is usually subject to a combination of these factors. Promising candidates are identified *via* an exhaustive evaluation of all library candidates. While this strategy is highly parallelizable and can enable remarkable experimentation throughputs, it is also highly resource-demanding for large libraries and requires numerous experiments. The search for promising candidates can, in principle, be streamlined if feedback collected from evaluations of some candidates is used to hypothesize about the properties of the remaining candidates. When assuming a deterministic structure-property relation for the considered material system, structural similarity could be leveraged to estimate the properties of materials candidates. This strategy can be implemented in a closed-loop process, where artificial decision-makers use previously collected measurements to select the most promising materials candidates to evaluate iteratively. Viable search strategies balance the expectation on the properties of untested library candidates with the uncertainty on this estimate. In cases where the physical structure-property relation for the library candidates is unknown, data-driven approaches

can be used to infer materials properties from previously measured examples. For the rapid identification of the most desired library candidates, data-driven models do not necessarily need to achieve high prediction accuracies. Instead, the data-driven tools need to provide robust modeling capacities in low-data scenarios to estimate both expected properties and uncertainties on these estimates reliably.

ML offers several strategies to enable the robust modeling and identification of unknown structure-property relations in closed-loop processes, including Bayesian optimization and active learning. Bayesian optimization targets the perhaps slightly narrower goal of global black-box optimization to identify parameter values within a given domain for which an unknown response function yields optimal values. Active learning, in contrast, describes a broader class of scenarios where learning algorithms can query the evaluation of specific parameter points to reach pre-defined goals. Bayesian optimization can, therefore, be applied directly in closed-loop processes that aim to maximize or minimize a physical property, while active learning strategies allow us to broaden the scope and the goal of the closed-loop process to scenarios where the desired outcome is more vaguely defined. Such goals can include the sparse exploration of a considered parameter space or the maximization of diversity in the responses of queried points. The capacity of these ML models to condition decisions on collected feedback enables more directed and guided searches for promising parameters which have the potential to significantly reduce the number of experiments required to make a discovery.

In this part, we introduce three novel ML algorithms for scientific discovery with autonomous strategies in closed-loop processes. Chapter 6 introduces PHOENICS, a data-driven approach for the rapid identification of optimal values for continuous process parameters, such as annealing temperatures in fabrication procedures or compositions for multi-component materials which are abundant factors to the design organic solar cells (OSCs) (see Chapter 12). PHOENICS is implemented as a Bayesian optimizer which operates well on low-data regimes. The search policy used by PHOENICS aims to avoid redundant experiments to reduce the number of closed-loop iterations. Contrary to existing methods, PHOENICS leverages concepts of Bayesian kernel density estimation (BKDE), which natively enable the parallelization of the search, such that multiple experiments can be proposed in one closed-loop iteration. Chapter 7 extends this optimization framework to categorical parameters, which could encode the choice of material constituents, molecular structures, or individual processing steps. This algorithm, which we called GRYFFIN, provides a competitive optimization framework for categorical parameters. Unlike existing methods, GRYFFIN can accelerate the search significantly by exploiting physical domain knowledge in the form of continuous physicochemical

descriptors associated with the individual categorical choices. A dynamic sensitivity analysis with respect to the supplied descriptors allows GRYFFIN to identify the relevance of the provided descriptors to estimate the measured responses. With this feature, GRYFFIN can evidence physical insights and inspire scientific understanding. While both PHOENICS and GRYFFIN target the optimization of single objectives, Chapter 8 introduces an *a priori* method for multi-objective optimization on closed-loop processes. This approach, referred to as CHIMERA, allows researchers to express preference information based on an importance hierarchy in the objectives. CHIMERA enables global optimization algorithms to identify parameter points that satisfy the stated objectives in the order of the importance hierarchy. To this end, CHIMERA requires trade-offs for each objective to determine accepted degradations based on which improvements on less important objectives can be realized. By supporting the declaration of dynamic, relative trade-offs with respect to the achievable range of objective values, CHIMERA can be applied in closed-loop processes without requiring detailed knowledge of the response surfaces. All algorithms introduced in this part are designed with the goal to be deployed in autonomous experimentation workflows and are made available on GitHub.^{78–80}

PART III: AUTONOMOUS EXPERIMENTATION FOR LIGHT-HARVESTING TECHNOLOGIES

The capacity of ML approaches to model causal relations governing physical systems outlined in Part II provides opportunities to digitize some aspects of the scientific method and streamline scientific discovery. In combination with automated experimentation equipment, data-driven tools can enable autonomous workflows for scientific discovery in the search for optimal process parameters for an experimental setup or to identify the best design choices which lead to the discovery of a functional molecule or advanced material. Self-driving laboratories can realize this vision for autonomous experimentation, where experiments are designed by data-driven tools and executed by robotic platforms without human intervention. Several prototypic designs for autonomous experimentation platforms have recently been reported and are summarized in recently published comprehensive reviews.^{81–84} In principle, the closed-loop approach can be implemented sequentially, such that ML tools and automated platforms are invoked one after another until one iteration completes and the cycle starts again. Yet, the tight integration of ML tools and automated platforms opens opportunities for much more interleaved and streamlined workflows. Such workflows pose challenges on the control software and require a flexible and adaptive orchestration of autonomous experimentation. Chapter 9 reviews existing technologies following this vision, highlights the im-

mediate challenges to the wide-spread deployment of self-driving laboratories and outlines future developments which further advance autonomous experimentation for scientific discovery. Based on these considerations, we introduce CHEMOS in chapter 10, a first-generation operating system for self-driving laboratories. CHEMOS provides the essential layers indispensable for autonomous experimentation, including versatile databases, human-robot interactions, and automated results analyses. At its core are a variety of ML-driven strategies for experiment planning such as those introduced in Part II and interfaces to a selection of off-the-shelf robotic hardware for automated experimentation in chemistry and materials science. As such, CHEMOS lowers the obstacles to the implementation of closed-loop approaches to scientific discovery. This chapter demonstrates several diverse proof-of-concept applications of CHEMOS, highlighting the ease-of-use in addition to the benefits of ML augmented closed-loop experimentation over standard approaches.

Finally, in Chapters 11 and 12, we demonstrate two self-driving laboratories for autonomous experimentation to discover advanced light-harvesting technologies. Chapter 11 introduces ADA, a fully autonomous platform capable of optimizing thin films materials used for energy conversion, storage, and conservation. Specifically, ADA targets the maximization of the hole mobility of organic hole-transport material (HTM) commonly used in perovskite solar cells (PSCs). ADA modulates the optical and electronic properties of thin-film materials by simultaneously modifying compositions and processing conditions. The closed-loop approach to thin-film discovery is implemented with CHEMOS (see Chapter 10) and process parameters are optimized with PHOENICS (see Chapter 6). The autonomous experimentation approach enables ADA to be configured for a wider range of doping levels than those typically explored in manual experiments. This broader experimentation scope leads to an unexpected scientific finding, where the parameter space revealed a region of enhanced thermal stability of the generated thin-films. The benefits of autonomous experimentation can, therefore, go beyond increased experimentation throughput at reduced cost and exhibit an enabling character based on which unexpected scientific discoveries can be made. Chapter 12 presents an autonomous platform for the discovery of photostable polymer blends for OSCs. The platform introduced in this chapter is capable of synthesizing photoactive polymer blends at specific composition ratios subject to the choices made by the data-driven experiment planner, and assesses their photostability *via* spectroscopic measurements under harsh environmental conditions. Similar to the self-driving laboratory presented in Chapter 11, this platform is orchestrated by CHEMOS and leveraged PHOENICS for the rapid identification of promising materials compositions. This chapter further compares the advantages and limitations of autonomous experimentation to high-throughput

experimentation (HTE) and manual approaches on a quantitative level. Our comparisons indicate that autonomous experimentation can enable scientific investigations at significantly lower materials consumptions while theoretically achieving experimentation throughputs comparable to HTE when parallelizing across applications.

2

Background

Greenhouse gas emissions arising from human activity have been identified as the dominant cause of observed global warming since the mid-20th century.⁸⁵ The continued emission of greenhouse gases such as carbon dioxide or methane is expected to induce long-lasting climatic changes which can increase the likelihood of severe, pervasive and irreversible impacts for people and ecosystems.⁸⁵ Transitioning into a low-carbon economy with substantially reduced greenhouse gas emissions could alleviate these imminent economical, environmental and societal challenges.⁸⁶ The transformation to a low-carbon economy requires the widespread deployment of clean energy technologies, along with possibilities for carbon utilization, storage and capture.⁷ The abundance of solar power, and the fact that plants have successfully developed photochemical processes to convert sunlight into chemical energy, provides the opportunity to use it as a massive clean energy resource.⁸⁷ In fact, the energy provided by the sun is expected to be sufficient to satisfy the worldwide energy consumption.⁸⁸ The key to viable artificial light-harvesting systems is the discovery of materials which operate at high power conversion efficiencies (PCEs) with long life times and low production costs. However, materials discovery is an inherently time-intensive and resource-consuming process with typical time scales of up to a decade of basic and applied research before a material candidate reaches the market.⁸⁹

A common strategy to the discovery of functional molecules or advanced materials for light-harvesting technologies consists in the testing of several candidates, and modulating the structures of tested candidates to drive their properties closer to the desired behavior. In this spirit, the identification of a promising molecule or material can be related to the mathematical task of global optimization, which offers algorithmic techniques to determine the element of a set of available options which best satisfies a selection criterion. Scientific hypotheses for such discovery problems speculate about the performance of individual candidates, while experiments are used to measure properties of interest and generally collect data about the probed candidates. Yet, experimental

investigations are usually resource-demanding and theoretical descriptions often balance accuracy and computational cost. Statistical models can bridge this gap by estimating the properties of interest from empirical evidence, collected experimentally, computationally or other means (see Sec. 1.3). Bayesian statistics in particular provides a rigorous framework to hypothesize about the properties of not yet tested materials candidates purely based on empirical evidence. If hypotheses are generated *via* data-driven approaches, the interplay of suggesting an experiment and executing the experiment can be implemented in a closed-loop workflow which can drive the discovery process without further human intervention. Autonomous experimentation platforms implement this data-driven strategy to scientific discovery by leveraging robotic hardware and machine learning (ML) methods, as we will discuss in more detail in Part III. These novel tools have recently been demonstrated to enable new avenues to the discovery of artificial light-harvesting technologies to thus catalyze the transition to a low-carbon economy (see Chapters 11 and 12).^{90,91} This chapter provides the background to the basic concepts of the emerging autonomous platforms for clean energy research. Sec. 2.1 overviews existing artificial light-harvesting concepts and promising technologies. In Sec. 2.2 we illustrate connections between the scientific method and Bayesian statistics in the context of scientific discovery. We further discuss strategies for algorithmic, data-driven experiment planning leveraging real-time feedback in a closed-loop process in Sec. 2.3. Finally, we highlight some striking historical and more recent developments of laboratory automation in Sec. 2.4.

2.1 LIGHT-HARVESTING AS AN APPROACH TO A CARBON FREE ECONOMY

Converting sunlight into chemical energy constitutes a key metabolic step for many biological organisms. This process, which is catalyzed by the photon-induced separation of electronic charges, provides inspiration for the design of artificial light-harvesting devices. For example, photovoltaic systems create electrical voltage and current upon photon absorption,¹⁵ excitonic networks (see Chapter 3) are developed for efficient excitation energy transport,⁹² and functional materials powered by sunlight facilitate carbon dioxide (CO₂) reduction⁹³ and water splitting.⁹⁴ While artificial solar energy conversion is on the rise, current technologies need to be advanced to expedite the transition to a low-carbon economy. Detailed mechanistic understanding and structural insights into the physiological processes in biological organisms to harvest sunlight could inspire the design of more efficient and more robust artificial light-harvesting devices.

2.1.1 NATURAL LIGHT-HARVESTING

Over millions of years,⁹⁵ photoautotrophs, notably cyanobacteria and plants, have developed efficient and robust strategies to achieve direct solar-to-fuel conversion, a process referred to as *photosynthesis*. In this process, the absorption of photons drives chemical reactions to convert water and carbon dioxide into carbohydrate molecules such as sugars. The primary steps of natural photosynthesis involve the creation of spatially separated electron-hole pairs upon photon absorption in the light-dependent reactions (see Fig. 2.1a). The resulting electric potential drives the oxidation of water to oxygen in the light-independent reactions. This water-splitting process is at the heart of the energetics of photosynthesis.⁸⁷ The key processes of the light-dependent reactions are facilitated by self-assembled light-harvesting pigment-protein complexes at high energy conversion efficiencies and robustness:¹⁵ the formation of electronic excitations induced by photon absorption as the primary energy conversion step, followed by excitation energy transfer (EET) and finally charge-separation, *via* charge-transfer excitations, to drive chemical reactions (see Fig. 2.1a).

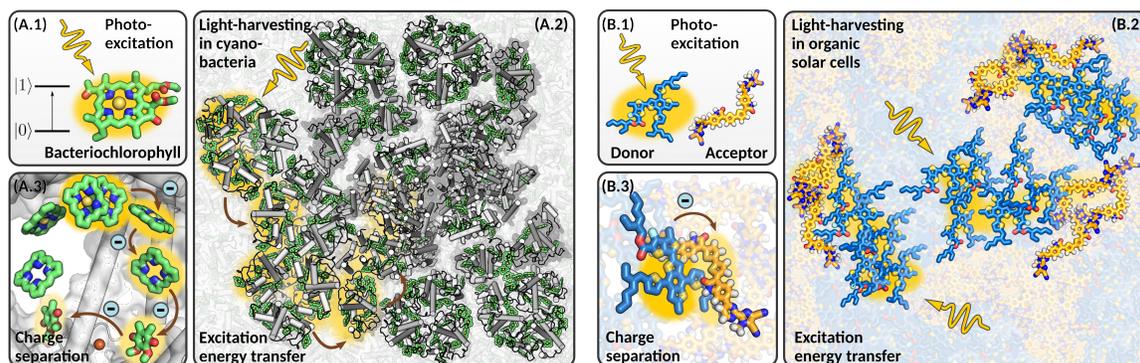


Figure 2.1: Light harvesting in phototrophic organisms (A) and organic solar cells (B). (A) Variants of the chlorophyll pigment molecules can create excitons upon photon absorption, which are transferred to the reaction center for charge separation. (B) Excitons created upon photon absorption by the donor material are transferred to the donor-acceptor interface for charge separation.

At optimal environmental conditions, photoautotrophs can convert almost all absorbed photons into stable photoproducts⁹⁶ and thus operate at nearly 100 % quantum efficiency.^{97,98} However, their energy conversion efficiency, quantified *via* the energy content of the harvestable biomass produced relative to the local annual solar irradiance, typically does not exceed 1 % for crop plants,^{99,100} or 3 % for microalgae.¹⁰¹ Artificial light-harvesting devices to date can achieve almost as high quantum efficiencies at overall higher energy conversion efficiencies.¹⁰² However, solar energy conversion efficiency must ultimately be assessed from the perspective of complete life cycles of the biologi-

cal organisms or the light-harvesting device.^{96,103,104} Photosynthetic organisms are more concerned about survival than high biomass production. Adverse, rapid changes in the incident photon flux are accounted for by small structural changes of one or more of the light harvesting proteins to open up energy dissipation pathways and thus limit the formation of harmful photoproducts such as reactive oxygen species.^{105,106} For this reason, the vital property of the photosynthetic apparatus is functional robustness in constantly fluctuating environments. Light-harvesting technologies based on low-cost molecular materials including polymers, organic semiconductors, and nanoparticles, face similar challenges of phototoxicity. The design principles on which natural photosynthesis operates could therefore inspire the design of more robust and long-lived light-harvesting materials.

2.1.2 THEORETICAL DESCRIPTIONS OF LIGHT-HARVESTING PROCESSES

Theoretical analyses and descriptions of natural photosynthesis rely on known microscopic structures of the involved pigment-protein complexes.¹⁰⁷ Green sulfur bacteria have emerged as one of the most widely used model organisms to study the mechanistic details of photosynthesis.^{108–126} Their light-harvesting apparatus is generally composed of photosynthetically active bacteriochlorophyll units which are spatially arranged into several interconnected functional units. The main element is the chlorosome, an ellipsoidal shaped body densely packed with bacteriochlorophyll-c pigments which, is generally regarded as an antenna complex to absorb incoming photons.¹²⁷ The chlorosome is connected to the baseplate¹²⁸ which funnels excitations to the reaction center (RC) to induce the charge separation (see Fig. 2.1). In all of these complexes, protein scaffolds promote a precise spatial arrangement of bacteriochlorophyll pigments. The first crystal structure of a light-harvesting pigment-protein complex was resolved in 1974, and named Fenna-Matthews-Olson (FMO) complex after the scientists who lead these efforts.^{129,130} The FMO complex funnels excitations to the RC and is commonly found in green sulfur bacteria. The next milestone in the structural analysis of photosynthesis was the resolution of the RC on the atomic level, as mentioned earlier in Chapter 1.

From photon capture to charge transfer, quantum mechanical phenomena are at the heart of the fundamental processes governing photosynthesis. Full theoretical descriptions and detailed understanding of these processes are most desirable to derive roadmaps for the design of artificial light-harvesting devices. Over the last decade, EET and charge-transfer events in large photosynthetic complexes have been a topic of interest from both a theoretical and experimental perspective.^{111,115,131,132} Obtaining the dynamics of the excitonic processes in photosynthesis requires to

solve the time dependent Schrödinger equation,

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = H |\psi(t)\rangle, \quad (2.1)$$

for the light-harvesting system and its environment, described by a wave function $|\psi(t)\rangle$ and a Hamiltonian H . However, only the two body problem has an exact analytic solution under the Schrödinger equation, such that the dynamics of larger systems can only be obtained numerically. The large system sizes, typically exceeding 100,000 atoms, and the relevant time scales ranging from ultrafast electronic processes in the order of femto- or picoseconds to slow reorganization events over micro- or milliseconds render a full quantum mechanical treatment of all degrees of freedom computationally intractable.¹³³ In fact, closing the gap between the length-scales of single molecules and macroscopic materials as well as the time-scales constitutes one of the outstanding computational challenges.^{41,134} Physical models for EET and charge transfer excitations in light-harvesting complexes typically describe the studied complex as an open quantum system. In this approach, the total Hamiltonian H of the light-harvesting complex is decomposed into contributions from a system S , a bath B , and interactions between the system and the bath,

$$H = H_S + H_B + H_{SB}. \quad (2.2)$$

The system S describes a subset of all degrees of freedom, usually the pigment molecules, while the remaining degrees of freedom, *i.e.*, the protein scaffold and the solvent, are summarized in the bath. The time evolution of the entire system can be obtained from the Liouville equation,

$$\frac{\partial}{\partial t} \rho(t) = -\frac{i}{\hbar} [H, \rho(t)], \quad (2.3)$$

where $\rho(t) = \sum_j |\psi_j(t)\rangle \langle \psi_j(t)|$ denotes the density matrix. The decomposition of the total Hamiltonian allows us to compute the system dynamics explicitly while treating the bath dynamics implicitly and can thus significantly lower the computational demand. Computational methods for open quantum system dynamics commonly leverage hybrid quantum mechanics/molecular mechanics (QM/MM) simulations, where the electronic structure of the system is modeled quantum mechanically, while the surrounding bath is described by a classical force field.¹³⁵

Molecular excitations in natural photosynthesis are governed by excitations between the ground and the first excited states, and are modulated by thermal fluctuations in the nuclear geometries

of the pigments and their surroundings. In a first approximation, excitation energy correlation functions can be determined from low-cost quantum chemistry methods for estimating excitation energies of molecular pigments in conformations generated with classical molecular dynamics.^{114,136,137} Extensions of this approach include quantum mechanical corrections to molecular geometries at increased computational demand,¹³⁸ or ground state dynamics calculations based on density functional theory (DFT).¹³⁹ Several methods for the numerical treatment of open quantum system dynamics have been developed. Numerically accurate methods such as the hierarchical equations of motion (HEOM) approach,^{119,140,141} or the hierarchy of pure states (HOPS)^{142,143} can treat system-bath interactions non-perturbatively and account for non-Markovian noise correlations but are computationally demanding, and can only be afforded for selected systems (see Chapter 3). Approximative schemes allow to compute the dynamics under certain assumptions, such as weak or strong couplings between the system and the environment at lower computational demand but might violate some physical constraints, for example the positivity of the density matrix or the correct thermalization of the final state. Breuer and Petruccione provide an comprehensive in-depth overview of the most common open quantum system dynamics approaches.¹⁴⁴

2.1.3 SELECTED SOLAR CELL MODELS

The bio-inspired design of molecules and materials for light-harvesting applications has been of interest for decades.¹⁴⁵ Solar cells constitute artificial devices which, inspired by the light-dependent reactions of photosynthesis, convert sunlight into electrical energy by generating spatially separated electron-hole pairs upon photon absorption. Several device architectures have been proposed, differing in their material constitutions and compositions. The National Renewable Energy Laboratory (NREL) maintains a comprehensive chart with existing and emerging solar cell technologies and their achieved efficiencies.¹⁴⁶ The efficiency of the light-to-energy conversion process is determined by the electronic properties of the constituting materials that regulate photon absorption and exciton dissociation events.

Solar cells can generally be categorized into three generations with individual mechanisms, advantages and drawbacks. Most of the commercially available solar cell technologies belong to the first generation, which are based on *pn*-junctions created from inorganic materials in the form of doped polycrystalline or single-crystal silicon.¹⁴⁶ First generation solar cells facilitate charge separation as a spontaneous process and to date achieve remarkable power conversion efficiencies, but require

purified silicon which poses a major cost to the fabrication process. Second generation solar cells are based on inorganic thin film materials including cadmium telluride (CdTe)^{147,148} and copper indium gallium selenide (CIGS) technologies.¹⁴⁹ Although this class of solar cells does not reach power conversion efficiencies as high as first generation solar cells, they can be fabricated at much lower cost. Emerging third generation photovoltaic cells offer a wide spectrum of low-cost applications, including inorganic and organic materials, which are actively being research as they have the potential to achieve higher power conversion efficiencies.

For example, emerging third-generation designs in the form of hybrid organic-inorganic perovskites promise to offer more cost effective and competitive solutions if produced at large scale,¹⁵⁰ although their efficiencies are currently below that of first- or second-generation devices.¹⁴⁶ perovskite solar cells (PSCs),^{151–153} are typically composed of inorganic lead halide matrices, and contain inorganic or organic cations. Power conversion in PSCs is achieved by the direct absorption and conversion of incoming photons into free electrons and holes. The free charges are then extracted through *p*- and *n*-type contacts. Recently, PSCs have experienced increased attention as breakthroughs in materials and device architectures boosted their efficiencies and stabilities.¹⁵⁴ The architectures in which PSCs can function efficiently can slightly differ in the role the perovskite material in the device and the nature of the electrodes. In Chapter 11, we report an autonomous platform for the discovery of optimal hole-transport materials (HTMs) for the p-type contact of PSC architectures.

In contrast, organic solar cells (OSCs) have been suggested as an alternative architecture for third-generation designs. OSCs use phase-separated mixtures of two or more organic materials in bulk-heterojunction architectures to absorb light and split the exciton into electron-hole pairs at the interface between the materials (see Fig. 2.1b).^{155–157} Thus, OSCs fall somewhere between the limits of natural photosynthesis and crystalline solar-cells with desirable properties often limited by energetic and structural disorder.^{15,158,159} In comparison to their inorganic counterparts, OSCs exhibit several appealing advantages such as mechanical flexibility or lower energy payback times. Initially, OSC designs have been proposed with fullerenes as acceptor materials due to their excellent electron-transporting properties and favorable bulk heterojunction morphology.^{160–163} Yet, fullerene-based OSCs are limited by fundamentally constrained energy levels,¹⁶⁴ and photochemical instability¹⁶⁵ which limits their efficiency to <12%.¹⁴⁶ The discovery of several families of non-fullerene acceptor molecules provided a major advance in engineering OSC candidates.^{166,167} Currently, these acceptors replace fullerene derivatives in all highly-efficient OSCs,^{156,157} reaching power conversion efficiencies well beyond the most efficient fullerene-based OSCs. The large number of degrees of freedom arising

from the complex aromatic structures allows to fine tune the electronic properties of non-fullerene acceptors including their optical gap, exciton diffusion length, exciton binding energy, the energy level alignment with the donor, or the charge-carrier mobility. Further development is required to make OSCs based on non-fullerene acceptors ready for commercial applications, mostly to reduce their chemical complexity and thus enable their inexpensive production on large scales. One approach to further improve OSC properties is demonstrated in Chapter 12, where we introduce an autonomous platform to enable the discovery of photostable multi-polymer blends.

2.2 SCIENTIFIC DISCOVERY WITH BAYESIAN STATISTICS

Statistics and statistical analyses are at the heart of investigations in the physical sciences and fundamental to quantifying the correspondence between the theoretical predictions of a physical model and experimentally observed responses and properties of the studied system (see Sec. 1.3).⁴⁶ Statistical methods are used to collect and organize empirical evidence, but can also make inferences under the framework of probability theory when modeling experiments as random events.¹⁶⁸ While probability has been axiomatically formalized as a real-valued measure ranging from 0 to 1 (see Sec. 2.2.1), two major competing interpretations of probability with different views about its fundamental nature have emerged.¹⁶⁹ An objective approach to probability is commonly described by the frequentist’s interpretation, which understands probability as the limit of relative frequencies or occurrences of an experiment after infinitely many trials.¹⁷⁰ In this interpretation, the probability of an event can be approximated with a finite number of trials by counting the number of times the event occurred, compared to the number of times the experiment was conducted. A probability of 0 indicates that the event does not occur, while a probability of 1 denotes that the event always occurs. Contrary subjective approaches to probability, however, such as Bayesian interpretations, state that probabilities express a degree of belief in the occurrence of an event.^{171,172} In this interpretation, the degree of belief can range from 0 (*disbelief*, or FALSE) to 1 (*certainty*, or TRUE). This interpretation has long-ranging implications to approaching scientific discovery *via* probabilistic statements, which we will explore in detail in Sec. 2.2.2

The appeal to apply a Bayesian interpretation of probability to experimental contexts is based on the fact that it does not require infinitely many repetitions of an experiment, allows to generalize events to propositions or statements, and can infer probabilities for an event from both expert knowledge and experimental data.¹⁷³ Bayesian statistics formulates statements about physical sys-

tems modeled by a set of unobservable variables or not yet observed quantities in terms of probability distributions. These distributions are conditioned on modeling decisions possibly based on domain knowledge and observed events obtained from empirical data. Since Bayesian statistics treats probabilities as a degree of belief, Bayesian inference can quantify distributions on the values of the modeling parameters and adjust them to the experimental evidence.¹⁷⁴ The possibility to condition statements and conclusions on observed data sets Bayesian inference apart from commonly described retrospective approaches to statistical inference, and allows to directly relate Bayesian data analysis to the scientific method (see Sec. 1.1), as we will explore in more detail in Sec. 2.2.3.

2.2.1 BRIEF INTRODUCTION TO PROBABILITY AXIOMS

Probability is a measure which is used to describe random events. To define probability axiomatically, we first introduce a few technical terms commonly used in probability theory. We consider *experiments* which are conducted under defined test conditions and can have different *outcomes*. The outcomes of an experiment are stochastic, which implies that the specific outcome of a single execution of the experiment can change every time the experiment is conducted. However, the set of all possible outcomes for an experiment is known in advance and summarized in the *sample space*. Finally, an *event* describes a set of outcomes of an experiment, and the *event space* summarizes all unique events for the experiment. Events which contain exactly one outcome are called *elementary events*. With these definitions, we can introduce probability as a measure space, (Ω, F, p) , where p represents the probability of an event E of the event space F , denoted as $p(E)$, and Ω comprises the sample space. The probability p measures the occurrence of an event E and follows three axioms, which were first formalized by Kolmogorov in 1933:^{175,176}

- (i) *Non-negativity*: the probability p of an event $E \subset F$ is a non-negative real number,

$$p(E) \in \mathbb{R}, p(E) \geq 0 \quad \forall E \in F. \quad (2.4)$$

- (ii) *Unitary*: the probability that at least one of the elementary events in the entire sample space will occur is 1,

$$p(\Omega) = 1. \quad (2.5)$$

(iii) σ -*additivity*: any countable sequence of disjoint sets of events, E_1, E_2, \dots , satisfies

$$p\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} p(E_i). \quad (2.6)$$

Following these three axioms, additional properties of probability as a measure can be derived.¹⁷⁷ For example, one can show that probability is a monotonic measure, which implies that the probability of an event A which is a subset of an event B is less than or equal to the probability of B , $p(A) \leq p(B)$. A consequence of this observation is that the probability of the empty set, *i.e.* the probability that no event occurs, is zero.

2.2.2 DERIVATION OF BAYES' THEOREM

Bayes' theorem, alternatively known as Bayes' rule, relates the probability of an event to prior assumptions about the event and data collected from empirical evidence about occurrences of the event. Inference in the Bayesian context invokes Bayes' theorem to update the degree of belief about an event, expressed as probability, based on additional data. To derive Bayes' theorem, we first introduce the notion of *conditional probability* for two events A and B . In some cases, the observation of an event B allows us to update our belief about the occurrence of an event A . This notion of conditioning is captured by the conditional probability of A given B , denoted as $p(A|B)$, which quantifies the belief about A knowing that B has already happened. We further consider the conjoint probability of the two events A and B . The probability of the event that both events A and B have occurred, denoted as $p(A \cap B)$, can be expressed for two possible scenarios: (i) event B could occur with probability $p(B)$, and subsequently trigger event A with probability $p(A|B)$, or (ii) event A could occur with probability $p(A)$, and subsequently trigger event B with probability $p(B|A)$. The probability $p(A \cap B)$ of both events occurring can thus be expressed as

$$p(A \cap B) = p(A|B)p(B), \quad (2.7)$$

$$= p(B|A)p(A), \quad (2.8)$$

where Eq. 2.7 corresponds to scenario (i) and Eq. 2.8 corresponds to scenario (ii). From these two equations, we can find an expression for the conditional probability of A given B ,

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}, \quad (2.9)$$

which already constitutes one formulation of Bayes' theorem. The predictive power of Bayes' theorem can be rationalized when assuming that some events can be observed and others cannot. For example, the beta radiation sources used by Kaufmann in the early 20th century (see Sec. 1.1) emit electrons upon the decay of specific radium isotopes. While the emission of an electron can be directly observed, the decay rate of the radium isotope is not directly observable. Yet, we can estimate the decay rate of the radium isotope indirectly by detecting and counting the emitted electrons. In the example discussed for the derivation, we can consider A to be an unobservable event while B describes an observable event. For this scenario, Eq. 2.9 infers information about the unobservable event A from observations of possibly several instances of B . We thus get access to information about A and can refine our belief about A indirectly *via* the related observable B . The distinction between observable and unobservable events might not always be precisely defined or possible at all, but illustrates the capacity of Bayes' theorem to statistically model phenomena of the physical world.

2.2.2.1 DIACHRONIC INTERPRETATION

The relevance of Bayes' theorem to the scientific method can be justified under the diachronic interpretation of Eq. 2.9. This interpretation allows us to consider statistical models describing physical phenomena, and refining these models based on empirical evidence collected, *e.g.*, in laboratory experiments or computer simulations (see Sec. 1.2). As such, the diachronic interpretation can be used to identify the statistical model which best describes a studied physical phenomenon or a structure-property relation in light-harvesting materials. With the Bayesian interpretation of probability as a degree of belief, A and B do not necessarily need to be interpreted strictly as events. Instead, we can consider A to represent a hypothesis, H , for a causal relation in a physical system, and B to summarize the data, D , that we have collected about the system in laboratory experiments or generally represent empirical evidence. For example, we empirically estimate and discuss the relation between optical and electronic properties of organic electronics materials in Chapter 4. With this

interpretation, Eq. 2.9 transforms to

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}. \quad (2.10)$$

Although Eq. 2.10 is mathematically identical to Eq. 2.9, this diachronic interpretation of Bayes' theorem clearly provides a rigorous approach to refining the belief about a given hypothesis, $p(H)$, given the belief about collected data $p(D)$ and the conditional probability to have observed this data under the hypothesis at question, $p(D|H)$. If the hypothesis H contains adjustable parameters, θ , this interpretation of Bayes' theorem provides a mathematical framework to refine these parameters based on collected data. One step further, this interpretation of Bayes' theorem can be used as a tool for the data-driven validation or rejection of scientific hypotheses. The individual terms of Eq. 2.10 are commonly interpreted as follows:¹⁷⁸

- (i) *Prior*, $p(H)$: The prior summarizes our belief about a hypothesis H before observing any data. As such, the prior includes all modeling decisions, background information, domain expertise and our physical expectations about the hypothesis, which could be inspired from experience, intuition or scientific concepts.
- (ii) *Likelihood*, $p(D|H)$: The likelihood describes the belief that the considered hypothesis H has generated the observed data D by quantifying the probability to have observed D under the assumption that H is true.
- (iii) *Evidence*, $p(D)$: The evidence accounts for our belief in the data, *i.e.*, it accounts for the probability of having observed the data under any valid hypothesis. Computing the evidence can be numerically involved or even entirely intractable such that approximative schemes need to be leveraged.
- (iv) *Posterior*, $p(H|D)$: The posterior constitutes a refined belief in our hypothesis H after having collected data D about the physical system.

The computation of the normalizing constant can be computationally intractable, especially for large hypothesis spaces. The calculations can be simplified if the hypothesis space can be constructed such that the hypotheses are mutually exclusive, *i.e.*, at most one hypothesis can be true, and collectively exhaustive, *i.e.*, at least one of the hypotheses is true. However, this simplification is not always applicable, such that Bayesian methods are typically regarded as computationally demanding compared to frequentist approaches to data analysis.

2.2.2.2 CHOOSING A PRIOR

The belief about a hypothesis H after collecting data D depends on the prior belief about the hypothesis, $p(H)$ (see Eq. 2.10). Thus, the choice of the prior directly affects the posterior belief. Choosing a prior is a non-trivial task and usually requires background knowledge about the studied system. Generally, the prior can be constructed following two different approaches: (i) the population approach, and (ii) the state of knowledge approach.¹⁷¹ The population approach interprets the prior distribution as a population of possible hypotheses from which the hypothesis of current interest is drawn. In this context, the prior represents the probability with which the considered hypothesis is formulated. The state of knowledge approach, in contrast, has a more subjective interpretation of the prior distribution in the absence of a population of hypotheses, and expresses our prior belief based on accumulated knowledge and uncertainty about this knowledge. Regardless of the interpretation, the prior distribution should comprise all plausible hypotheses such that the hypothesis space is collectively exhaustive, but does not necessarily need to be localized around the true value assuming the possibility to accumulate sufficient amounts of data to eventually outbalance the choice of the prior and dominate the shape of the posterior (see Sec. 2.2.2.3).

In some cases, where no background information is available, a sensible choice might consist in the construction of an uninformative prior. Such a choice can be justified via the principle of insufficient reason, *i.e.*, the lack of any prior knowledge, or the expectation that all outcomes are equally likely. The construction of an uninformative prior, however, is not straightforward and requires careful considerations. For many tasks, there is no clear choice of a flat prior distribution. A distribution which is flat in one parametrization of the hypothesis space might not be flat if the space is reparameterized. This observation illustrates the difficulty of applying the principle of insufficient reason, as it is not always clear on what scale it should apply. One approach to alleviate this challenge consists in the construction of Jeffrey's prior,¹⁷⁹ which is invariant under coordinate transformations on the parameters of a hypothesis. This invariance implies that relative probabilities assigned to volumes of the probability space are independent of the parametrization of the problem. Jeffrey's prior is constructed from the square root of the determinant of the Fisher information matrix. In any case, the search for a truly uninformative prior and prioritizing a particular specification of unformativity might not be necessary for practical applications as the likelihood is expected to dominate the posterior with sufficient empirical evidence, such that minor differences in generally flat priors will not significantly impact the posterior.

2.2.2.3 STATEMENTS ABOUT THE POSTERIOR

Bayesian inference describes the process of computing a posterior distribution from a prior distribution following Bayes' theorem (see Eq. 2.10). Assuming that the hypothesis H is parametrized by parameters θ , and that data D are collected in the form of samples y , Bayes' theorem can be used to express some general relations between the prior $p(\theta)$ and the posterior $p(\theta|y)$. For example, we can relate the mean of the posterior distribution to the mean of the prior distribution and find¹⁷¹

$$E(\theta) = E(E(\theta|y)). \quad (2.11)$$

This equation states that the prior mean of θ is identical to the average of all possible posterior means over the distribution of possible data y . For the variances of the prior and the posterior we find¹⁷¹

$$\text{var}(\theta) = E(\text{var}(\theta|y)) + \text{var}(E(\theta|y)), \quad (2.12)$$

which states that the expected posterior variance $E(\text{var}(\theta|y))$ is smaller than the prior variance $\text{var}(\theta)$. This observation has long ranging implications, as it suggests that the collection of observations and the incorporation of this acquired information into the posterior generally reduces our uncertainty in the belief about the hypothesis. Thus, Bayesian inference immediately relates to knowledge acquisition in the scientific discovery context (see Sec. 1.1). Yet, the two relations in Eqs. 2.11 and 2.12 only describe expectations, such that the posterior variance can be similar to or even larger than the prior variance in some cases. An increase in the variance with the collection of data, however, might indicate a conflict or inconsistency between the sampling model and the prior distribution.

2.2.3 BAYESIAN DATA ANALYSIS

Data analysis is a process ubiquitous to the physical sciences, the life sciences and the social sciences, which covers the inspection, transformation and modeling of empirical data with the goal to extract useful information, to draw conclusions and to support decisions.¹⁷¹ Although data analysis is a collective term which covers multiple facets and approaches, we will mostly focus on data analysis in a Bayesian context in this section and explicitly use probability to quantify the uncertainty and

degree of our belief on the derived inferences. Bayes' theorem as derived in Sec. 2.2.2 provides the foundation for this approach as it describes a rigorous framework for statistical inference, *i.e.*, to draw conclusions based on data that is subject to random variation such as observation errors and sampling variations (see Sec. 1.2). Bayesian data analysis includes the following three steps:

- (i) *Modeling*: The modeling step does not require empirical data yet. Instead, this step focuses on the construction of a complete probabilistic model to describe all observable and hidden variables of the considered physical system. This probabilistic model needs to be constructed such that it spans a joint probability distribution for all involved variables. In addition, the probabilistic model needs to be consistent with the physical knowledge about the underlying scientific problem. This aspect includes the adherence to conservation laws and other physical constraints. Once constructed, the probabilistic model represents the prior to the inference step and should thus include all background knowledge and expectations on the behavior of the studied system.
- (ii) *Inference*: The inference step combines the constructed prior with collected empirical data into the posterior *via* Bayes' theorem (see Eq. 2.9). Depending on the complexity of the constructed model, this step can be computationally involved or even intractable. For this reason, considerations about the feasibility of the inference step are sometimes accounted for already when making modeling decisions such that the prior probabilistic model is approximated, or possible shapes of the posterior distribution are constrained to known, computationally tractable probability distributions.
- (iii) *Reasoning*: The final step of Bayesian data analysis involves the detailed evaluation and analysis of the posterior model. This process could involve assessments of the predictive power of the constructed model, *i.e.*, determining how accurately the constructed model fits the data, and if the observed data can be explained with the inferred model at all. Assuming the model generally constitutes a good fit, predictions about observed and hidden physical parameters of the model can be made, which could inspire conclusions or decisions about the studied physical system. The reasoning step is usually subjective as acceptable model accuracies can vary from task to task, and specialized analyses might be required to reason about particular problems.

The steps of Bayesian data analysis can thus be directly related to the steps of the scientific method which are outlined in Sec. 1.1 by associating the probabilistic model constructed in the modeling step

with the testable hypothesis in the scientific method, and interpreting the inference and reasoning steps as the collection of evidence from experiments and the refinement of the hypothesis based on the empirical evidence. To illustrate the process of Bayesian data analysis, we consider an example which is inspired by Chapter 8 of *Think Bayes* by Allen Downey.¹⁷⁸

M2 ANALYSIS

The Boston area is a thriving center for scientific research with lots of opportunities to learn. One such opportunity is a seminar series called *TheoChem*, which is jointly hosted between Boston University, the Massachusetts Institute of Technology (MIT), and Harvard University. *TheoChem* is designed for and run by graduate students at these three institutions who invite researchers in the fields of theoretical chemistry and physics to give lectures on their favorite research topics. Topics cover diverse areas spanning open quantum systems dynamics and network biology. Overall, this seminar series is a great opportunity for graduate students to meet with the leading scientists in their fields and learn about cutting edge research. However, most of the lectures are hosted at MIT, which requires graduate students from Harvard and Boston University to travel to MIT to attend the seminars.

Luckily, Harvard operates shuttle buses, such as the M2, which takes affiliates safely from the Harvard main campus to the medical campus in downtown Boston and stops at MIT in between. Although the M2 shuttle operates on a fixed schedule* the traffic in the Cambridge and Boston area can cause irregularities in the schedule of the otherwise reliable M2 shuttle. Given the distance between the Harvard and the MIT campuses and the expected travel times by bus, it could sometimes be faster to walk to MIT, while at other times it might be worth to wait for the next M2 shuttle. We know that if there are many people waiting at the bus stop, the M2 could arrive shortly, but we might have to wait for a significantly longer time if there are only a few people at the bus stop by the time we get there. The problem we would like to address in this example concerns the question of finding the fastest option to go from Harvard to MIT, walking or taking the bus, given the number of people waiting at the bus stop. For a Bayesian analysis of this task, we first start with some modeling decisions.

PREMISE 1: We model the time interval, τ , between two consecutive M2 shuttle buses as a random variable, where the expected time interval between two consecutive shuttles follows a

* The weekday schedule of the M2 shuttle bus can be found here:
https://www.masco.org/system/files/downloads/m2_vanderbilt_weekdays.pdf

pre-determined schedule but fluctuations in the time interval arise from varying traffic conditions. *TheoChem* seminars are usually scheduled for 4 pm, when the M2 operates on a targeted interval of 10 minutes. We assume that the time interval between two consecutive shuttles, which we refer to as *gap*, can be modeled as a gamma process described by a constant shape parameter, α , and a constant inverse scale parameter, β ,

$$\tau \sim \text{Gamma}(\alpha, \beta), \quad \text{where } p_{\text{gap}}(\tau|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}, \quad (2.13)$$

where $p_{\text{gap}}(\tau|\alpha, \beta)$ denotes the probability distribution for the time between two consecutive busses, τ , given a shape of α and an inverse scale of β . The distribution of gaps, p_{gap} , is illustrated in Fig. 2.2b for $\alpha = 10$ and $\beta = 1 \text{ min}^{-1}$, which matches the expected time interval of 10 minutes.

PREMISE 2: The number of passengers, n , arriving at the bus stop within a certain time, t , can be modeled as a random variable. Further, passengers arrive at a constant rate, λ , and the arrival of any passenger is independent from the arrival of any other passenger, such that we can model the arrival of passengers at the bus stop with a Poisson process,

$$n \sim \text{Poisson}(\lambda, t), \quad \text{where } p_{\text{passengers}}(n|\lambda, t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad (2.14)$$

where $p_{\text{passengers}}(n|\lambda, t)$ denotes the probability distribution for observing n passengers at the bus stop who arrived at a constant rate λ after a time interval t has passed. The distributions of passengers, $p_{\text{passengers}}$, is illustrated in Fig. 2.2a.

PREMISE 3: Given the irregularities in the M2 schedule caused by traffic, we do not check the schedule or any shuttle trackers when we leave for the bus stop, such that we can arrive at any point during a gap of duration τ between two consecutive shuttles.

These premises conclude our initial modeling process. However, we will see later on that further modeling decisions are required, as it is typically the case in Bayesian data analysis where choices during individual steps might affect choices made in other steps. Before we transition to the inference step, we will first consider a subtle implication of our model regarding the probability distribution of the gaps and the probability to observe a gap of duration τ . Although the gaps τ between two consecutive shuttles follow a Gamma process expressed *via* $p_{\text{gap}}(\tau|\alpha, \beta)$ as stated in PREMISE 1,

this distribution does not resemble the distribution of gaps that we will observe. In fact, given our arbitrary arrival time as stated in PREMISE 3 we are more likely to arrive during a longer gap than during a shorter gap. The distribution of observed gaps will thus be skewed towards longer gaps. This expectation reflects a phenomenon known as *observer bias*.¹⁷⁸ To compute the distribution of observed gaps, $p_{\text{gap,obs}}(\tau|\alpha, \beta)$, we need to weight the probability of each gap, $p_{\text{gap}}(\tau|\alpha, \beta)$, by its duration τ and normalize the resulting distribution across all possible gaps,

$$p_{\text{gap,obs}}(\tau|\alpha, \beta) = \frac{\tau p_{\text{gap}}(\tau|\alpha, \beta)}{\int_0^\infty d\tau \tau p_{\text{gap}}(\tau|\alpha, \beta)}. \quad (2.15)$$

The normalizing constant is formulated via an integral over all possible values of the gap, and can be evaluated *via* integration by parts

$$\int_0^\infty d\tau \tau p_{\text{gap}}(\tau|\alpha, \beta) = \int_0^\infty d\tau \tau \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} = \int_0^\infty d\tau \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^\alpha e^{-\beta\tau} \quad (2.16)$$

$$= -\frac{\beta^{\alpha-1}}{\Gamma(\alpha)} \tau^\alpha e^{-\beta\tau} \Big|_0^\infty - \int_0^\infty d\tau \frac{-\beta^{\alpha-1}}{\Gamma(\alpha)} \alpha \tau^{\alpha-1} e^{-\beta\tau} \quad (2.17)$$

$$= 0 + \frac{\alpha}{\beta} \int_0^\infty d\tau \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} = \frac{\alpha}{\beta}, \quad (2.18)$$

which matches the expected duration of a gap. The distribution of observed gaps, $p_{\text{gap,obs}}(\tau|\alpha, \beta)$, can therefore be expressed as

$$p_{\text{gap,obs}}(\tau|\alpha, \beta) = \frac{\beta}{\alpha} \tau p_{\text{gap}}(\tau; \alpha, \beta) \quad (2.19)$$

$$= \frac{\beta}{\alpha} \tau \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \quad (2.20)$$

$$= \frac{\beta^{\alpha+1}}{\alpha\Gamma(\alpha)} \tau^\alpha e^{-\beta\tau} = \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} \tau^{(\alpha+1)-1} e^{-\beta\tau} = p_{\text{gap}}(\tau|\alpha+1, \beta). \quad (2.21)$$

The distribution of observed gaps thus simplifies to the same functional form as the distribution of gaps, but with an updated value of the shape parameter of $\alpha+1$, which induces an overall shift to longer gaps in the distribution. This observation is one of the consequences of the observer bias, as unlikely events are undersampled. Fig. 2.2b illustrates the distribution of observed gaps, $p_{\text{gap,obs}}$ in comparison to the distribution of gaps, p_{gaps} .

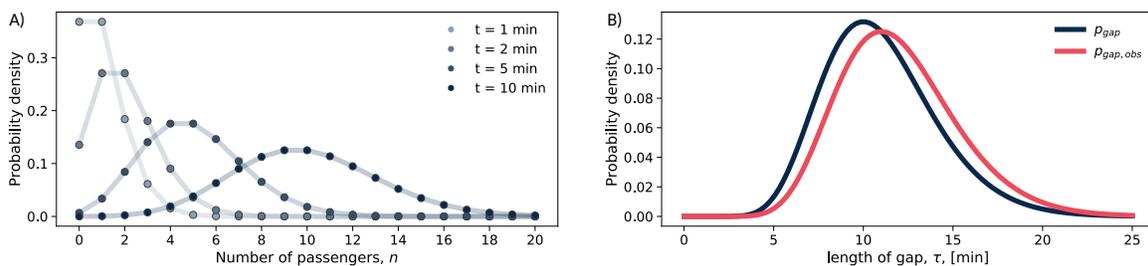


Figure 2.2: Distributions of the governing processes of the M2 problem. (A) Distributions of passengers waiting at the bus stop depending for different times since the last bus arrived. The number of arrived passengers is modeled as a Poisson process with a constant rate of 1 passenger/minute. (B) Distribution of time intervals between two consecutive busses (gaps), and the distribution of gaps as observed by a passenger. Gaps are modeled with a Gamma process with a shape of $\alpha = 10$ and an inverse scale of $\beta = 1 \text{ min}^{-1}$.

UNINFORMED WAITING TIMES

We start our analysis by evaluating the predictive power of our initial model solely based on the gaps between two consecutive shuttle busses and without accounting for the passengers who are waiting at the bus stop by the time that we arrive there. In this first analysis, we will derive a probability distribution for the wait time t , which we define as the time between our arrival and the arrival of the next shuttle bus. We can relate our waiting time to the length of the gap τ during which we arrive *via*

$$\tau = t_0 + t, \quad (2.22)$$

where t_0 denotes the time between the previous bus and our arrival. Since we do not check the bus schedule (see Premise 3), we can assume that we arrive at an arbitrary time t_0 during a randomly drawn gap of length τ . We can account for this assumption by modeling the arrival distribution $p_{\text{arrival}}(t_0)$ as a uniform distribution on the interval $[0, \tau]$, *i.e.*,

$$t_0 \sim \text{Uniform}(0, \tau), \quad \text{where } p_{\text{arrival}}(t_0|\tau) = \frac{\Theta(\tau - t_0)\Theta(t_0)}{\tau}, \quad (2.23)$$

where $\Theta(\cdot)$ denotes the Heaviside step function. With the analytic form of the arrival distribution p_{arrival} in Eq. 2.23 and the possibility to express the arrival time as $t_0 = \tau - t$ (see Eq. 2.22) we find that the distribution of the wait time, $p_{\text{wait}}(t)$, is also a uniform distribution, specifically

$$t \sim \text{Uniform}(0, \tau), \quad \text{where } p_{\text{wait}}(t|\tau) = \frac{\Theta(t)\Theta(\tau - t)}{\tau}. \quad (2.24)$$

This expression for the distribution of the wait time is still conditioned on the length of the gap, τ , during which we arrive. To determine the unconditional distribution of the wait time, we need to marginalize over all possible gaps and weight the conditional wait times based on the probability to observe a gap of length τ . It is important to note that our example requires the observation of a gap of length τ , which is why the use of $p_{\text{gap,obs}}$ instead of p_{gap} is more appropriate here. The distribution of the wait time can thus be calculated as

$$p_{\text{wait}}(t) = \int_0^{\infty} d\tau p_{\text{wait}}(t|\tau) p_{\text{gap, obs}}(\tau|\alpha, \beta) \quad (2.25)$$

$$= \int_0^{\infty} d\tau \frac{\Theta(t)\Theta(\tau-t)}{\tau} \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} \tau^{\alpha} e^{-\beta\tau}, \quad (2.26)$$

$$= \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)} \int_t^{\infty} d\tau \tau^{\alpha-1} e^{-\beta\tau}, \quad (2.27)$$

which cannot be solved in closed form. This distribution constitutes the prior predictive distribution, which we can use to make predictions without any data-driven refinement solely from the prior probabilistic model. Fig. 2.3 illustrates the probability density and the cumulative probability of this distribution. We observe that the distribution of the wait time is almost flat with generally high values for wait times below 5 minutes, and gradually decreases for larger wait times. The decrease in probability for larger wait times can be attributed to the fact that observing large gaps τ well beyond the expected gap length of 10 minutes is unlikely. From the prior predictive we could infer if, on average, waiting for the next shuttle bus take us to MIT faster, assuming that our modeling decisions are reasonable. In the following, we will also account for empirical evidence.

INFORMED WAITING TIMES

We can refine our wait time analysis by including the number of people waiting at the bus station by the time of our arrival. Previously, we assumed that our arrival time since the last shuttle bus, t_0 , followed a uniform distribution. Now, however, we want to condition the belief about our arrival time on the number of people waiting at the bus stop. While we cannot directly measure our arrival time, given that the arrival time of the previous shuttle bus is unknown, we can measure the number of people at the bus stop and use this evidence to update our belief about our arrival time using Bayes' theorem. Given that passengers arrive at a constant rate λ (see PREMISE 2), we would generally expect t_0 to be larger if there are more passengers waiting for the shuttle. For a

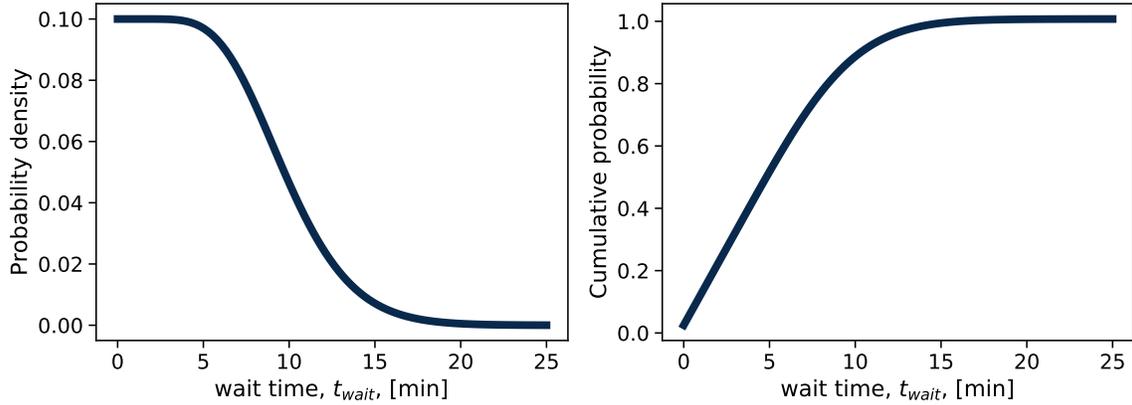


Figure 2.3: Distribution of wait times when ignoring the number of passengers waiting at the bus stop. The illustrations are generated for $\lambda = 1$ passenger/minute, $\alpha = 10$ and $\beta = 1 \text{ min}^{-1}$. (A) Probability density function of the wait time. (B) Cumulative density function of the wait time.

quantitative analysis, we consider a case where we arrive during a gap of length τ and observe n passengers waiting for the shuttle. We are interested in computing the distribution of arrival times, $p_{\text{arrival}}(t_0)$ given n passengers who arrived at rate λ during a gap of length τ , and can calculate this distribution with Bayes' theorem

$$p_{\text{arrival}}(t_0|n, \lambda, \tau) = \frac{p_{\text{passenger}}(n|\lambda, t_0, \tau)p_{\text{arrival}}(t_0|\tau)}{p_{\text{passenger}}(n|\lambda, \tau)}. \quad (2.28)$$

In this equation, $p_{\text{passenger}}(n|\lambda, t_0, \tau)$ denotes the likelihood and as such quantifies the probability to observe n passengers during a gap of length τ after a time t_0 has passed, $p_{\text{arrival}}(t_0|\tau)$ is the prior and represents the probability to arrive at time t_0 during a gap of length τ , and $p_{\text{passenger}}(n|\lambda, \tau)$ is the empirical evidence, which constitutes the probability to observe n passengers during a gap of length τ . With our modeling assumptions, we can derive the posterior of the arrival time

$$\begin{aligned} p_{\text{arrival}}(t_0|n, \lambda, \tau) &= \frac{p_{\text{passenger}}(n|\lambda, t_0, \tau)p_{\text{arrival}}(t_0|\tau)}{p_{\text{passenger}}(n|\lambda, \tau)} \\ &= \frac{\frac{(\lambda t_0)^n e^{-\lambda t_0}}{n!} \frac{\Theta(t_0)\Theta(\tau-t_0)}{\tau}}{\int_0^\infty dt_0 \frac{(\lambda t_0)^n e^{-\lambda t_0}}{n!} \frac{\Theta(t_0)\Theta(\tau-t_0)}{\tau}} \\ &= \Theta(t_0)\Theta(\tau-t_0) \frac{t_0^n e^{-\lambda t_0}}{\int_0^\tau t^n e^{-\lambda t}}. \end{aligned} \quad (2.29)$$

This updated distribution for the arrival time accounts for both our prior assumption that we do not check the bus schedule and the observed number of passengers at the bus stop by the time that

we arrive. From this posterior, we can calculate a posterior wait time distribution similar to the uninformed case using $t_0 = \tau - t$ from Eq. 2.22.

$$p_{\text{wait}}(t|n, \lambda, \tau) = \Theta(\tau - t)\Theta(t) \frac{(\tau - t)^n e^{-\lambda(\tau - t)}}{\int_0^\tau dt t^n e^{-\lambda t}}, \quad (2.30)$$

which expresses the distribution of waiting times within a given gap of length τ . Since we do not know the length of the gap during which we arrive, we need to further compute the posterior predictive distribution for the wait time, $p_{\text{wait}}(t|n, \lambda)$, by marginalizing over all possible gaps,

$$\begin{aligned} p_{\text{wait}}(t|n, \lambda) &= \int_0^\infty d\tau p_{\text{wait}}(t|n, \lambda, \tau) p_{\text{gap,obs}}(\tau|\alpha, \beta) \\ &= \int_0^\infty d\tau \Theta(\tau - t)\Theta(t) \frac{(\tau - t)^n e^{-\lambda(\tau - t)}}{\int_0^\tau dt t^n e^{-\lambda t}} \frac{\beta^{\alpha+1}}{\Gamma(\alpha + 1)} \tau^\alpha e^{-\beta\tau} \\ &= \frac{\beta^{\alpha+1}}{\Gamma(\alpha + 1)} \int_t^\infty d\tau \tau^\alpha e^{-\beta\tau} \frac{(\tau - t)^n e^{-\lambda(\tau - t)}}{\int_0^\tau dt t^n e^{-\lambda t}}. \end{aligned} \quad (2.31)$$

The posterior predictive distribution, $p_{\text{wait}}(t|n, \lambda)$, is illustrated in Fig. 2.4 for different numbers of passengers waiting at the bus stop at arrival. We find that the posterior predictive distributions assume shapes which are inherently different from the prior predictive distribution due to the fact that the number of passengers waiting at the bus stop at arrival informs us about the time which has passed since the last bus arrived, regardless of how many passengers are waiting when we arrive. Assuming a travel time of a little more than 15 minutes when taking the M2 from Harvard Yard to the *TheoChem* seminar in MIT's Building 4, and a walking time of about 25 minutes, our analysis indicates that we will likely arrive at MIT sooner by foot if there are fewer than five people waiting at the bus stop by the time we get there. However, we made some fairly strong assumptions in our derivation. For example, rate of passengers arriving at the bus stop is probably not constant, as passengers tend to check the schedule and Harvard's shuttle tracker is a frequently used tool. Passengers are also unlikely to arrive indendently, as they could form groups to share the time of their commute with one another. We can probe the sensitivity of our results on these modeling choices by slightly modyfing the modeling choices and repeating the analyses. Alternatively, we can construct a more accurate model which avoids these assumptions by monitoring the passenger behavior for a given period of time to record the true distribution of passenger arrivals.

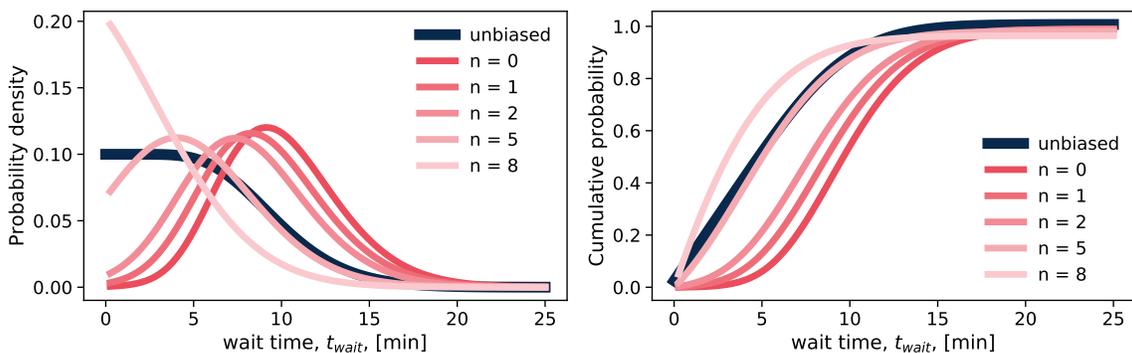


Figure 2.4: Distribution of wait times when including the number of passengers waiting at the bus stop. The illustrations are generated for $\lambda = 1$ passenger/minute, $\alpha = 10$ and $\beta = 1 \text{ min}^{-1}$. (A) Probability density function of the wait time for different numbers of passengers waiting at the bus stop at arrival. (B) Cumulative density function of the wait time for different numbers of passengers waiting at the bus stop at arrival.

2.3 EXPERIMENT PLANNING IN THE BAYESIAN CONTEXT

Experiments require the preparation of an initial state for the studied physical system conditioned on controllable parameters which modulate the responses or the properties of the system (see Sec. 1.2). For example, the geometric arrangement of molecular pigments can change the efficiency of EET across the set of pigments (see Chapter 8) and the polymer composition of a multi-component OSC can drastically alter its photostability (see Chapter 12). To formulate experiment planning as an optimization task, we consider n controllable parameters on a defined domain, $\mathbf{z} \in \mathcal{Z}^n$, which could be continuous, discrete or categorical variables, and an experimental response, $f: \mathcal{Z}^n \rightarrow \mathbb{R}$, for each of the parameter choices within the domain, which we assume to be a scalar for now. Generalizations to the simultaneous optimization of multiple properties will be discussed in Sec. 2.3.2. The experimental response could be obtained from a single measurement or multiple measurements, and generally represents the merit assigned the parameters to express how well the evaluated candidate satisfies the desired targets.

When optimizing synthesis protocols for organic donor or acceptor molecules in OSC, the controllable parameters could for example include the reaction temperature, the amount of solvent, and the choice of the catalyst, while the experimental response could be quantified *via* the rate at which the desired product is produced. We assume that the selection criterion which assesses the merit of a given set of parameter values is already included in the scalar response f . As such, the optimization task breaks down to the identification of the specific parameter values, $\mathbf{z}^* \in \mathcal{Z}^n$, which yield the most desired experimental outcome, $f(\mathbf{z}^*)$. For simplicity, we will consider the optimization task as

a minimization task from hereon, *i.e.*, we formulate f such that $\mathbf{z}^* = \underset{\mathbf{z} \in \mathcal{Z}}{\operatorname{argmin}} f(\mathbf{z})$ corresponds to the desirable experimental outcome. To identify this optimal set of parameter values, optimization strategies can in principle leverage feedback from previously conducted experiments. Every time we evaluate a set of parameter values, \mathbf{z}_k , we can record the associated response, $f_k = f(\mathbf{z}_k)$, to gradually collect a set of observations, $\mathcal{D}_n = \{\mathbf{z}_k, f_k\}_{k=1}^n$, which can guide our experiment planning strategies to select promising candidates.

Compared to other optimization tasks in science and engineering, such as finding energy minima in many-body systems or the optimization of the shape of an aircraft wing, where the response can usually be determined computationally, the identification of optimal parameters for laboratory experiments poses a few additional challenges due to the nature of the experiment. In the following, we list a few assumptions about laboratory experiments in chemistry and materials science which set them apart from other optimization tasks.

- (i) We consider moderately large parameter spaces for the optimization task, *i.e.*, the parameters $z \in \mathcal{Z}^n$ are defined on domains with moderately large values of n . For most laboratory experiments, we can assume that $n \leq 20$.
- (ii) The parameter domain on which we optimize is bounded and the geometry of the parameter domain is known, such that it is straightforward to inexpensively assess the membership of an arbitrary set of parameter values to the optimization domain. In most cases, the parameter domain will correspond to a multi-dimensional cuboid.
- (iii) The experimental response f constitutes a surface which lacks a known special structure like concavity or linearity. We thus need to assume that f can adopt any shape and consider f as a black-box response.
- (iv) When measuring the response f , we can only measure the response itself and no first or higher order derivatives are experimentally accessible.
- (v) The experimental response f is subject to noise, *i.e.*, whenever we evaluate f , we observe the true function value with additional systematic and stochastic noise of unknown source and magnitude. Yet, f is not dominated by noise.
- (vi) Neglecting the noise, the response surface f is otherwise smooth almost everywhere, with the exception of isolated jumps or discontinuities.

(vii) Evaluations of f are expensive in terms of budgeted resources. Consequently, the number of evaluations that can be afforded is limited, and rarely exceeds a few hundred. Budget limitations can arise due to monetary expenses related to purchasing laboratory equipment or consumables, time requirements due to long experimentation steps or opportunity costs, for example when experimenting with human subjects or rare materials.

A plethora of optimization algorithms for the identification of global optima has been developed to solve problems in these contexts. Straightforward approaches to global optimization rely on the exhaustive evaluation of a large set of candidate solutions without leveraging any feedback collected from response evaluations. Although such strategies do not implement any policies to direct the search, they can be quite efficient in certain applications, especially in high-dimensional spaces, and provide the compelling advantage of being massively parallelizable, which is demonstrated in high-throughput (HT) approaches to experimentation.^{180–183} Exhaustive searches can follow randomized,^{184–186} quasi-randomized,¹⁸⁷ or systematic grid searches and (fractional) factorial designs.^{188–190} Such approaches have, for example, been demonstrated in the context of chemical reactions,^{180,181,191,192} biomedical research,¹⁸² and the discovery of methane storage materials¹⁸³ and electrolytes.¹⁹³ Gradient-based algorithms such as gradient descent,¹⁹⁴ conjugate gradient,¹⁹⁵ or quasi-Newtonian methods,^{196,197} are highly efficient at optimizing convex surfaces in the absence of noise. However, most experimental surfaces are expected to be non-convex and although the response mostly dominates the noise, experimental feedback is still influenced by measurement uncertainties, which renders this class of optimization algorithms mostly inapplicable. Isolated examples of gradient-based algorithms for experiment planning have been reported for the optimization of organic reactions.¹⁹⁸ Genetic algorithms and evolutionary strategies^{199–201} extend the idea of a random exploration of the search space, but base their exploration policies on a population of candidate solutions which have already been evaluated. New candidates are generated based on random perturbations of the candidate population. During the optimization, better candidates substitute poorly performing candidates based on different heuristic criteria.²⁰² Genetic strategies have recently been reported for the controlled growth of carbon nanotubes,²⁰³ the optimization of nanoalloy clusters,²⁰⁴ or the measurement of electronic spectra of rotamers of organic compounds.²⁰⁵ Examples of such methods include the covariance matrix adaptation evolution strategy (CMA-ES), which samples candidates from a multinomial distribution on the parameter space.^{206,207} After evaluating all proposed parameter points, distribution parameters are updated via a maximum-likelihood approach.

Other examples include particle swarms optimization,^{208,209} which defines a set of equations of motion to propagate candidate solutions in the parameter space. Random perturbations of candidate solutions in combination with heuristic criteria to accept these perturbations have also been implemented in physics-inspired methods such as simulated annealing,^{210,211} or tabu searches.^{212–214} Recently, Bayesian optimization^{215–220} has gained increased attention as a competitive global optimization strategy for various applications,^{221,222} including automatic ML,^{223–225} engineering design,^{226,227} and experimental design.^{228–231} Bayesian optimization describes a class of data-driven gradient-free optimization strategies which are designed for tasks budgeted optimization tasks in noisy environments.²³² In this section, we will review the methodology of Bayesian optimization as a competitive experiment planning strategy in more detail and further describe approaches to the simultaneous optimization of multiple objectives, following the background discussions of recently published studies.^{233–235}

2.3.1 BAYESIAN OPTIMIZATION FOR SINGLE OBJECTIVES

Bayesian optimization constitutes a gradient-free approach to global optimization tasks where evaluations of the response function f are resource demanding and only a budgeted number of evaluations can be afforded.²³² It has been shown that Bayesian optimization can locate global optima with very few function evaluations compared to other optimization strategies by reducing redundancies in the proposed candidates.²²² However, Bayesian optimization methods are typically much more computationally demanding than alternative approaches, and are thus only applicable to moderately large parameter domains with at most a few tens of parameters to be optimized simultaneously. This restriction, however, fits well with our assumptions about laboratory experiments. The common framework of a Bayesian optimization strategy follows two basic steps: (i) the construction of a surrogate as a data-driven approximation to the unknown response surface from a probabilistic ML model based on collected measurements, and (ii) the selection of new candidates with an acquisition function which balances the expected performance of each candidate and the uncertainty on this estimate as determined by the surrogate. The surrogate is usually determined by refining a probabilistic model *via* Bayesian inference on the collected data (see Sec. 2.2.2). To this end, the surrogate model is constructed from a prior distribution, $\phi_{\text{prior}}(\theta)$, over functions, possibly described by parameters θ . The surrogate is intended to approximate the response surface, while evaluations of the surrogate are expected to be much cheaper than evaluations of the true surface. During the

optimization, we collect pairs of evaluated candidates \mathbf{z}_k and associated responses $f_k = f(\mathbf{z}_k)$ into a pool of n observations, $\mathcal{D}_n = \{(\mathbf{z}_k, f_k)\}_{k=1}^n$. These observations are used to compute a posterior $\phi_{\text{post}}(\boldsymbol{\theta})$ for the parameters of the surrogate model with the goal to approximate the unknown response function. The surrogate is usually required to converge to the response surface in the limit of infinitely many distinct function evaluations. Several choices for priors and probabilistic models have been suggested, which are briefly summarized in Sec. 2.3.1.1. In the second step of the general Bayesian optimization procedure, the surrogate is used to identify the most promising candidate for future evaluation. Greedy search strategies could suggest the candidate which yields the optimal response under the surrogate. However, the use of probabilistic models to construct the surrogate allows to balance the expected response with the uncertainty about this estimate. More rapidly converging search strategies leverage the uncertainty of the surrogate to balance the exploitation of acquired observations with the exploration of the space to overall identify the global optimum at a faster pace. Different policies have been introduced, which we will review in Sec. 2.3.1.2.

2.3.1.1 CONSTRUCTING SURROGATES TO THE OBJECTIVE FUNCTION

Constructing the surrogate to an unknown optimization problem is a challenging task. Without any prior knowledge about the shape of the response surface, the surrogate needs to have the flexibility to model any type of function. In addition, the surrogate needs to not only estimate the expected response for an unseen candidate, but also provide an uncertainty on this estimate. For these reasons, surrogates are constructed from non-parametric probabilistic models. A frequently used statistical model to construct the surrogate is the Gaussian process (GP).^{220,236–239} Gaussian processes associate every point in the parameter domain with a normally distributed random variable. These normal distributions are constructed *via* a similarity measure between individual observations given by a user-defined kernel function. A GP provides a flexible analytic approximation to the response surface. Yet, inferences on a GP are computationally involved as they require the inversion of a possibly dense covariance matrix and thus scale cubically in the number of observations. With this limitation, Gaussian processes are typically only used in relatively low dimensional optimization tasks where the optimum can be located with relatively few function evaluations. Another choice for probabilistic models to construct the surrogate are random forests (RFs).^{240–244} They are constructed as an ensemble learning model from a collection of regression trees, and have been shown to perform particularly well for discrete input data and classification tasks. RFs are therefore success-

fully applied to objective functions with discrete or quasi-discrete co-domain. The computational cost of inference and prediction with RFs is much more favorable than with Gaussian processes due to a linearithmic scaling with the number of observations and a linear scaling with the dimensionality of the parameter space. Yet, RFs tend to overfit their training data, and uncertainty estimates can only be obtained empirically.²⁴⁵ Recently, Bayesian neural networks (BNNs) have been suggested for Bayesian optimization,^{246,247} retaining the flexibility of Gaussian processes at a favorable computational scaling comparable to RFs. In contrast to traditional neural networks, weights and biases for neurons in BNNs are not single numbers but instead sampled from distributions. BNNs are trained by updating the distributions from which weights and biases are sampled *via* Bayesian inference (see Sec. 2.2.2). The use of BNNs in Bayesian optimization has further inspired hybrid approaches, where the surrogate is constructed implicitly via Bayesian kernel density estimation (BKDE), for example in the recently introduced PHOENICS and GRYFFIN algorithms (see Chapters 6 and 7).^{233,234}

2.3.1.2 CONSTRUCTING ACQUISITION FUNCTIONS FROM THE SURROGATES

With the construction of a surrogate from a probabilistic model, we can define an acquisition function which balances the exploitation of collected data with the exploration of the parameter space based on the estimates and uncertainties of the surrogate. While purely exploitative strategies can be successful on convex response surface, the explorative component of the acquisition function is required to overcome local optima and quickly identify the global optimum. The ideal acquisition function finds the adequate balance between these two strategies. The exploration of the parameter space should be favored when no observations in vicinity to the global optimum have been made yet and the acquisition function should only sample close to the global optimum once its general location has been determined. One of the earliest and most widely applied acquisition functions is *expected improvement* and variants thereof.^{218,227,237} Expected improvement aims to measure the expected amount by which an observation of a point in the parameter space improves over the current best value. Exploration and exploitation are implicitly balanced based on the posterior mean and the estimated uncertainty. More recently, alternative formulations of acquisition functions have been developed. The *upper confidence bound* method constructs an acquisition function based on confidence bounds.^{248,249} Variants of this acquisition function have been designed specifically to be applied in higher dimensional parameter spaces.^{238,250} *Predictive entropy* estimates the negative differential entropy of the location of the global optimum given the observations,^{251,252} and has been

shown to outperform expected improvement and upper confidence bound acquisitions. In Chapter 6 we further introduce a novel acquisition function which alleviates some of the limitations of the aforementioned examples and natively enables batch-wise optimization *via* an intuitive sampling parameter.

2.3.2 OPTIMIZING MULTIPLE OBJECTIVES AT ONCE

Designing a novel light-harvesting device is a challenging decision-making process which involves numerous design choices tuning device parameters to improve several properties at once. In the case of OSCs or PSCs, for example, both the device efficiency and stability are typically of interest (see Sec. 2.1). Economically viable solutions, however, also require an inexpensive assembly of the device from the composing materials and that the composing materials are abundantly available. Multi-objective (Pareto) optimization targets the simultaneous optimization of a set of objective functions, $\{f_j\}_{j=1}^m$, where each of the objective functions, f_j , is defined on the same compact parameter space \mathcal{Z}^n .²⁵³ Objectives of interest in the context of chemistry could be, for example, the yield of a reaction and its execution time. Although the desired goal of an optimization procedure is to find a point in the parameter space $\mathbf{z}^* \in \mathcal{Z}^n$ for which each of the objectives $f_j(\mathbf{z}^*)$ assume their desired optimal value. Yet, objectives in multi-objective optimization tasks oftentimes conflict each other. For example, the functionalization of an acceptor candidate in OSCs might improve its electronic properties, but likely also complicate its chemical synthesis and shorter execution times of chemical reactions could cause a drop in yield. Improving on one objective could therefore imply an unavoidable degradation on other objectives. As a consequence, a single global solution cannot be defined for the generic multi-objective optimization task.

2.3.2.1 DEFINING AND IDENTIFYING SOLUTIONS TO MULTI-OBJECTIVE OPTIMIZATION PROBLEMS

A commonly used criterion for determining solutions to multi-objective optimization problems is *Pareto optimality*.²⁵⁴ A point is called Pareto optimal *if and only if* there exists no other point such that all objectives are improved simultaneously. Therefore, deviating from a Pareto optimal point always implies a degradation in at least one of the objectives. Relating to the previous example, this corresponds to a scenario in which the execution time cannot be improved any further without a degradation of the reaction yield. As Pareto optimal points cannot be collectively improved in

two or more objectives, solving a multi-objective optimization problem translates to finding Pareto optimal points. Note, that for a given multi-objective optimization task, multiple Pareto optimal points can coexist.²⁵⁵ Approaches to solving multi-objective optimization problems aim to assist a decision maker in identifying the favored solution from the set of Pareto optimal solutions (Pareto front). The favored solution is determined from preference information regarding the objectives. Methods for multi-objective optimization can be split into two major classes. *A posteriori* methods aim to discover the entire Pareto front, such that preferences regarding the objectives can be expressed knowing which objective values are achievable. *A priori* methods instead require preference information prior to starting the optimization procedure. As such, *a priori* methods can be more specifically targeted towards the desired goal and thus reduce the number of response evaluations if reasonable preference information is provided.

A posteriori methods are commonly realized as mathematical programming approaches such as *normal boundary intersection*,^{256,257} *normal constraint*,^{258,259} or *successive Pareto optimization*,²⁶⁰ which repeat algorithms for finding Pareto optimal solutions. Another strategy consists in evolutionary algorithms such as the *non-dominated sorting genetic algorithm-II*,²⁶¹ or the *sub-population algorithm based on novelty*,²⁶² where a single run of the algorithm produces a set of Pareto optimal solutions. Recently, *a posteriori* methods have also been developed following Bayesian approaches for optimization.^{263–267} However, determining the preferred Pareto point from the entire Pareto front requires a substantial number of objective function evaluations compared to scenarios in which only a subset of the Pareto front is of interest. Such scenarios can be found in the context of experiment design, where preferences regarding objectives like yield and execution time are available prior to the optimization procedure. As such, *a priori* methods appear to be better suited for multi-objective optimization in the context of autonomous experimentation, as they keep the number of objective evaluations to a minimum. A common *a priori* approach for expressing preferences for multi-objective optimization is to formulate a single cumulative function from a combination of the set of objectives which accounts for the expressed preferences. For example, instead of considering the yield and the execution time of a reaction independently, a single objective can be constructed from a combination of simultaneous observations for the yield and the execution time. Such cumulative functions are referred to as achievement-scalarizing functions (ASFs). The premise of the constructed ASF is that its optimal solution coincides with the preferred Pareto optimal solution of the multi-objective optimization task. Typically, ASFs are constructed with a set of parameters which account for the expressed preferences regarding the individual objectives. ASFs can be con-

structured *via, e.g.*, weighted sums or weighted products of the objectives. In such approaches, the ASF is computed by summing up each objective function f_k multiplied by a pre-defined weight w_k accounting for the user preferences. Multiple formulations of weighted sums and products exist,²⁶⁸ and methods have been developed to learn these weights adaptively.²⁶⁹ Weighted approaches are usually simple to implement, but the challenge lies in finding suitable weight vectors to yield Pareto optimal solutions. In addition, Pareto optimal solutions might not be found for non-convex objective spaces. A second *a priori* approach consists in considering only one of the objectives for optimization while constraining the other objectives based on user preferences.^{270–272} These approaches, referred to as ϵ -constraint methods, have been shown to find Pareto optimal points even on non-convex objective spaces.^{255,273} However, the constraint vector needs to be chosen carefully, which typically requires detailed prior knowledge about the objectives. A third *a priori* approach, known as lexicographic methods, follows yet a different approach.²⁷⁴ Lexicographic methods require preference information expressed in terms of an importance hierarchy in the objectives. In our example, when optimizing for the yield of a reaction and its execution time, the focus could be either on the reaction yield or on the execution time. In the scenario where the reaction yield matters the most, it is related to a higher hierarchy than the execution time. To start the optimization procedure with a lexicographic method, the objectives are sorted in descending order of importance. Each objective is then subsequently optimized without degrading higher-level objectives.²⁷⁵ Variants of the lexicographic approach allow for minimal violations of the imposed constraints.^{276,277}

2.4 LABORATORY AUTOMATION

Automation, the process of designing artificial systems capable to execute pre-defined actions without requiring further input, is commonly attempted to transfer repetitive tasks from human workers to mechanical or electrical machines. Laboratory automation specifically comprises technologies which are intended to enable the execution of laboratory processes for basic and applied research in the physical sciences and the life sciences with reduced or minimal human assistance. While laboratory robotics, focusing on the development of mechanical machines to execute pre-defined tasks, constitutes one of the largest subfields laboratory automation, the field more generally also includes the conception of non-standard experimentation processes and software algorithms to control and orchestrate robotic hardware. Automation technologies are usually introduced into laboratory environments to liberate the scientific workforce undesired labor, which for example includes repetitive

tasks or work in hazardous environments. Importantly, automated platforms can also enable protocols and procedures which cannot be executed by humans, for example monitoring experimental setups around the clock, measuring minute quantities of matter or sensing changes in the electromagnetic properties of a physical system. As such, laboratory automation technologies aim to support researchers in any and every step of an experiment, from setup over execution to measurement, and even provide end-to-end solutions for unsupervised experimentation without human intervention (see Part III).

However, introducing automation technologies into laboratory environments is a challenging endeavor. While automation in the life sciences has flourished ever since the introduction of the first automated DNA²⁷⁸ and peptide synthesizers,^{279,280} laboratory procedures in chemistry and materials science have proven to be much more difficult to automate. In fact, the relative homogeneity of biological molecules such as nucleic acids and proteins, the universality of water as the exclusive solvent, the emphasis on detection rather than synthesis, and the need to run thousands or millions of experiments in parallel, have all driven the transition towards experimentation with highly integrated systems in the life sciences. Common chemical operations, however, lack these advantages. Historical attempts and successes have emphasized crucial aspects to consider when using automated equipment in the laboratory to ensure that the set goals are met. Sec. 2.4.1 highlights some striking historical examples of laboratory automation leading towards autonomous experimentation to date, while Sec. 2.4.2 comments on the necessary considerations to take into account when automating laboratory processes.

2.4.1 HISTORICAL OVERVIEW OF LABORATORY AUTOMATION

Laboratory automation has a rich history, which can be dated back to the early efforts in mechanization and feedback control in Hellenistic technologies.²⁸¹ Water clocks, force pumps, float valves and weight regulators are only a few examples of these early automated devices, which were mostly based on hydraulic, pneumatic or mechanic mechanisms. First examples more targeted for applications in chemistry laboratories were reported in the form of automatic temperature regulators by Cornelis Drebbel in the early 17th century.^{282,283} Centrifugal (flyball) governors were conceived in the 18th century to regulate the moving speed of laboratory components or control valves.²⁸⁴ James Watt adapted this technology to control steam engines,²⁸⁵ and modifications of these devices to clock drives for telescopes and chronographs were developed by Throughton and Simms shortly after.²⁸²

Despite their simplicity and the lack of electric control components, centrifugal governors are still in use in common laboratory environments to date.²⁸⁶⁻²⁸⁸ While these examples constitute automated devices which could be used in laboratory environments, the design of automated machines specifically for laboratory applications can be dated back to the late 19th century.²⁸⁹ One of the earliest mentions is an automated washing apparatus reported in 1875 which can wash a filtrate by dripping water through a piece of filter paper at a defined rate.²⁹⁰⁻²⁹⁴ In the early 20th century, several similar devices had been introduced for the automated measurement of carbon dioxide in flue gasses to optimize combustion control.^{295,296} Although this apparatus was generally considered to reduce the workload of the operator, a significant amount of training and expertise was required to operate these devices, and only automated isolated steps of an experimentation process.

First end-to-end automated systems were introduced into research environments of the life sciences in the mid 1960s when Merrifield *et al.* reported an automated peptide synthesizer,^{279,280} a platform which fully automated all steps involved in the synthesis of peptides. This platform also constitutes one of the early examples of an experiment algorithmically controlled by a computer, which were more and more integrated into laboratory environments at that time.²⁹⁷ The IBM 1800 constitutes another milestone for time-shared laboratory automation systems and was used to control experiments and process collected measurement data. The system was connected to several laboratory instruments, including spectrometers,^{298,299} chromatographs³⁰⁰ and interferometer,³⁰¹ which could be used simultaneously and programmed independently. While the IBM 1800 provided automated solutions to individual experimentation steps, a more versatile reprogrammable robotic arm was introduced into the laboratory in 1984.³⁰² Controlled by a 16-bit microcomputer, the arm was able to carry out various operations to move laboratory equipment, such that operators could program the arm to execute more complex experimental tasks.

The widespread introduction of automation and end-to-end robotic platforms into laboratory environments was further intensified and streamlined by the industrial sector in an effort to increase the productivity and improve quality on chemical processes.³⁰³⁻³⁰⁹ The adoption of automation has also been driven in the medicinal context,³¹⁰⁻³¹³ and the pharmaceutical industry,³¹⁴⁻³¹⁷ where the introduction of the first automated DNA²⁷⁸ and peptide synthesis^{279,280,318} presented major breakthroughs in the field. During the last decade, fully automated systems have been reported for a diverse set of applications covering the synthesis of small organic molecules,³¹⁹⁻³²¹ the rapid screening of polymers,^{322,323} real-time monitoring of chemical reactions unraveling the kinetics of homogeneous and heterogeneous reactions,^{324,325} the identification of novel wide bandgap perovskites with higher

stability,³²⁶ and the discovery of colloidal nanocrystals,³²⁷ indicating that laboratory automation is now common to a large number of branches of science. More recent examples of end-to-end solutions for laboratory procedures with integrated platforms along with their advantages and shortcomings are discussed in Part III.

2.4.2 ADVANCING SCIENCE WITH AUTOMATED PLATFORMS

The promise of augmenting laboratory processes with automation to liberate the scientific workforce repetitive or challenging tasks and thus spark their creativity and eventually their productivity is overall appealing. Yet, automating a laboratory process is a demanding endeavor resulting in development costs which ultimately need to be measured in terms of both time investments and financial expenses. Most laboratory processes include some level of complication which poses a challenge to their automation. Often, these processes include individual steps which have not been systematically analyzed on a quantitative level and are not openly reported but instead passed down through generations of researchers as best practices.³²⁸ Introducing automated equipment into laboratory environments could also directly compromise the creativity and innovativeness of researchers. The perceived benefits from automation could motivate researchers to prioritize simple workflows which are more amenable to automation over more complex, manual procedures. Automating a laboratory procedure might also require adjustments to the precision, flexibility or the throughput at which experiments can be conducted. As such, scientific discovery is not immediately advanced only by using robotic hardware in the laboratory to perform tasks which could otherwise be done manually.³²⁹ Instead, innovation can be driven with deliberate and focused automation which enables researchers to solve a scientific question,²⁹⁷ rather than by defining a question which can be answered with automation.

Rewarding and enabling cases of laboratory automation can still target a variety of use cases. There are routine operations in chemistry laboratories, including weighing and labeling vials, which are part of many experimentation workflows.³²⁹ While automating these procedures will not drive innovation directly, the labor savings could still outweigh the development costs by sheer volume if these procedures are abundant to the present laboratory environment. Other applications might concern the execution of certain experiments at a large scale, for example selected chemical reactions to discover viable drug candidates. An HT system could be the key technology to upscale the experimentation process in this case, but focusing exclusively on the reaction implemented by the

HT system might shut down promising chemical spaces which can only be reached with other types of reactions. Frequently, the way that a human performs a procedure will not necessarily be the easiest way to perform the task in an automated fashion. Humans and robotic platforms can use different cues to collect information about an experimental procedure: humans can be sensitive to visual or auditory cues, while robotic platforms in principle can revert to electromagnetic cues. The most efficient approach to automating an existing laboratory procedure might therefore not consist in copying all human steps one by one. Instead, it could be more beneficial to rely on an automation engineer to first master the procedure itself, to then leverage the advantages and capabilities of robotic platforms to redesign the procedure from scratch, targeting the integrated implementation of an automated platform. Thus, laboratory automation can indeed eventually enable researchers to be more creative and innovative, but also only provides tools which need to be used correctly and with the appropriate purpose.

Part I

Machine learning in the sciences

This page is intentionally left blank.

3

Machine learning for quantum dynamics: deep learning of excitation energy transfer properties.

Apart from minor modifications, this chapter was originally published by the Royal Society of Chemistry as:

Machine learning for quantum dynamics: deep learning of excitation energy transfer properties. Florian Häse, Christoph Kreisbeck and Alán Aspuru-Guzik. *Chem. Sci.* **8** (12), 8419–8426 (2017).

Reproduced from Ref. [330] with permission from the Royal Society of Chemistry.

ABSTRACT

Understanding the relation between the structure of light-harvesting systems and their excitation energy transfer properties is of fundamental importance to several clean energy technologies, including the development of next-generation photovoltaics. Natural light-harvesting in photosynthesis shows remarkable excitation energy transfer characteristics, which suggests that pigment-protein complexes could serve as blueprints for the design of nature-inspired devices. Mechanistic insights into energy transport dynamics can be gained by leveraging numerically involved propagation schemes such as the hierarchical equations of motion. Solving these equations, however, is computationally costly due to the adverse scaling with the number of pigments. Virtual high-throughput screening, which has become a powerful tool in materials discovery, is, therefore, less readily applicable to the search of novel excitonic devices. In this chapter, we propose the use of deep neural networks to bypass the computational limitations of established techniques for exploring the structure-dynamics

relation in excitonic systems. Once trained, our neural networks reduce computational costs by several orders of magnitudes. The predicted transfer times and transfer efficiencies exhibit similar or even higher accuracies than frequently used approximative methods such as the secular Redfield theory. This study, therefore, presents an example of how data-driven approaches can lower the obstacles and the cost to testing structure-dynamics hypotheses for the discovery of novel excitonic devices for light-harvesting applications.

3.1 TRADITIONAL APPROACHES TO EXCITON DYNAMICS CALCULATIONS

Studying excitation energy transfer (EET) has been of great interest across several fields, bridging evolutionary biology to solar cell engineering for many years. Especially natural light-harvesting in the form of photosynthesis has been the subject of intense research (see Sec. 2.1). Pigment-protein complexes in plants and cyanobacteria exhibit remarkable transport properties which facilitate highly efficient EET across long distances.^{331–334} Identifying the governing principles of photosynthesis and ultimately transforming them into blueprints for novel nature-inspired excitonic devices is an active research frontier.^{335,336} Mechanistic investigations reveal valuable insights into the microscopic details of EET. Prominent studies probe the impact of electronic coherence or non-trivial interactions between excitons and specific vibrational modes on transfer characteristics.^{139,337–342} However such investigations are challenging since they require sophisticated experimental setups,^{115,340–344} or computationally involved accurate simulations of open-quantum system dynamics.^{139,143,337,338,345–347} Further, there are only a few fundamentally different natural light-harvesting complexes from which alone we cannot fully resolve the complex relation between the structure of an excitonic system and its dynamics in all detail.

In order to relate the dynamics to the underlying structure, it is desirable to investigate a large number of artificially designed excitonic systems. Such approaches have been addressed recently in several theoretical works.^{348–351} For example, the large-scale analysis of perturbations on pigment geometries in the Fenna-Matthews-Olson (FMO) complex revealed that higher transport efficiencies tend to be realized by more compact structures.³⁵² The

drawback of these empirical approaches is that they need to execute exciton dynamics calculations for a vast amount of randomly generated physically-plausible multi-chromophoric structures to collect sufficient empirical evidence. Due to the sheer number of performed dynamics simulations, such analyses quickly become computationally exhaustive, even if only the limits of physical approximations are studied with less demanding methods such as Lindblad equations.³⁵² The lack of readily available approaches to probe the structure-dynamics relation for excitonic systems, therefore, poses a significant bottleneck to the development of next-generation light-harvesting devices.

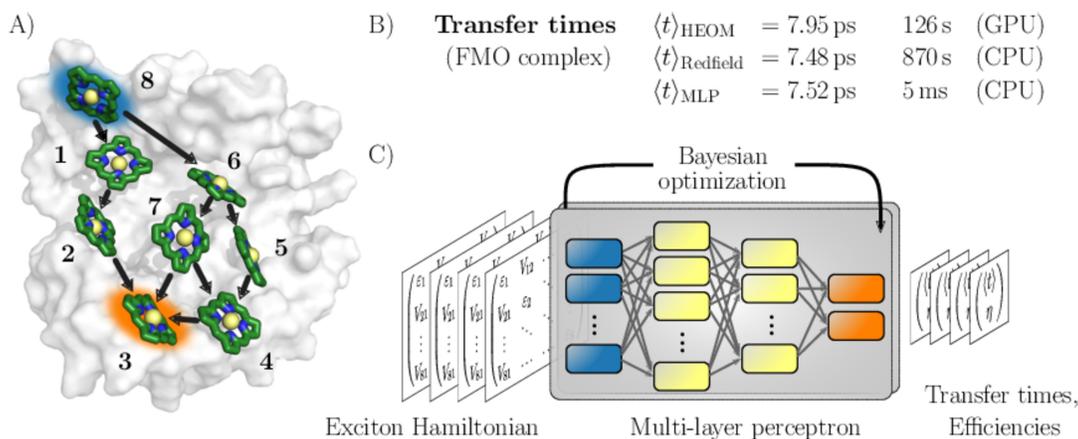


Figure 3.1: Machine learning excitation energy transfer properties in open quantum systems. (A) Fenna-Matthews-Olson (FMO) pigment-protein complex with eight chlorophyll pigments in the conventional numbering scheme. Dominant energy transfer pathways from the donor pigment 8 (blue) to the acceptor pigment 3 (orange) are indicated. (B) Results for average transfer time $\langle t \rangle$ calculations for energy transfer in the FMO complex from the donor to the acceptor obtained from solving the hierarchical equations of motion (HEOM), the approximate secular Redfield formalism and predicted by multi-layer perceptrons (MLPs) designed in this study. We report computational costs for each method. (C) Illustration of the MLP architecture. MLPs accept Frenkel exciton Hamiltonians as input features and predict average transfer times and efficiencies. The best network architectures were obtained through Bayesian optimization. Reproduced from Ref. [330] with permission from the Royal Society of Chemistry.

We suggest to follow a new path and leverage data-driven concepts from deep learning to bypass the computational demand of established techniques to explore EET properties (see Fig. 3.1). Specifically, we train multi-layer perceptrons (MLPs), a class of fully connected feed-forward artificial neural networks to predict average exciton transfer times and overall transfer efficiencies for a set of artificial excitonic systems. The input features to the MLPs are constructed from the parameters of the corresponding Frenkel exciton Hamilto-

nians (see Eq. 3.1).^{353,354} For the large scale screening of parameter spaces, the dynamics of only a fraction of all systems needs to be calculated to train the MLPs. Once trained, our neural networks evaluate transfer times within just a few milliseconds and thus bypass the computational demand of established techniques for exploring EET properties, while maintaining sufficiently high prediction accuracies. We demonstrate the potential of the MLPs on various artificial datasets which were generated by uniformly sampling pigment excitation energies and inter-pigment couplings in the vicinity of the energies and couplings of a several biologically relevant complexes: the FMO complex,¹²⁹ the light-harvesting complexes CP43,³⁵⁵ CP47³⁵⁶ and the reaction center (RC) of photosystem II.³⁵⁷ We aim to predict average transfer times from an initially excited donor to a certain acceptor pigment. Fig. 3.1 illustrates this prediction task for the FMO complex, which serves as an energy wire bridging the chlorosome and the reaction center in the photosynthetic apparatus of green sulfur bacteria (see Sec. 2.1).³⁵⁸ The initial excitation is assumed to be located at the donor pigment 8 since this pigment is in the proximity of the light-harvesting chlorosome antenna. The excitation energy then needs to be transferred to the target pigment 3, which couples to the RC where photochemical reactions are triggered. In the context of EET, the latter process is typically modeled as irreversible energy trapping.^{121,359–361}

The MLP models are trained based on transfer properties obtained with the hierarchical equations of motion (HEOM) formalism,^{119,140,141} which is a non-perturbative open quantum system approach taking into account non-Markovian effects. HEOM has become one of the standard tools in the field and serves as the ground truth in this study. The accuracy of the predictions critically depends on the choice of hyperparameters, such as the number of neurons, number of hidden layers, or the learning rate, which collectively define the specific architecture of the neural network. However, the best set of these parameters is *a priori* unknown. We determine the most suitable architectures for our MLP models from Bayesian optimization (see Sec. 2.3) on selected hyperparameters. This procedure is well-established in the machine learning (ML) community and was shown to outperform architectures built by domain experts.^{223–225,237} We assess the quality of the MLP predictions by comparing the

relative error of the predicted transfer times to the relative error made by secular Redfield calculations. The latter is simple to implement and commonly used to avoid the numerical complexity of more accurate HEOM simulations. Our findings suggest that MLPs provide a computationally cheaper alternative to secular Redfield calculations at comparable or, in most of our examples, even higher accuracy. As such, data-driven approaches have the potential to substitute resource-demanding parameter evaluations and can significantly accelerate the rate at which different excitonic systems are tested, for example, in closed-loop processes for autonomous discovery. Results for the FMO complex are summarized in Fig. 3.1.

3.2 MODELING OF EXCITATION ENERGY TRANSFER

The energy transport in light-harvesting complexes is determined by coupled molecular pigments which are embedded in a protein scaffold.^{362,363} The large number of degrees of freedom in the system renders a fully quantum mechanical treatment infeasible. Therefore, the exciton transfer dynamics is typically modeled with an effective Frenkel exciton Hamiltonian.^{353,354} The exciton Hamiltonian for a system of n sites for the single exciton manifold reads

$$H_{\text{system}} = \sum_{i=1}^n \epsilon_i |i\rangle\langle i| + \sum_{i \neq j}^n V_{ij} |i\rangle\langle j|, \quad (3.1)$$

where ϵ_i denotes the energy of the first excited state of the i -th pigment and V_{ij} denotes the Coulomb coupling between excited states at the i -th and j -th pigment. We assume that the excitonic system couples linearly to the vibrational environment of each pigment, which is modeled as a set of harmonic oscillators. The phonon mode dependent interaction strength is captured by the spectral density

$$J_i(\omega) = \pi \sum_k \hbar^2 \omega_{i,k}^2 d_{i,k}^2 \delta(\omega - \omega_{i,k}). \quad (3.2)$$

Here, $d_{i,k}$ defines the coupling strength of the k -th phonon mode ($b_{i,k}^\dagger$) of the i -th pigment with frequency $\hbar\omega_{i,k}$. In the first step of photosynthesis, energy is absorbed in the antenna pigments and subsequently transferred to the RC where photochemical reactions are triggered. This process can be described by energy transfer from an initially excited pigment (donor) to a target state (acceptor). We model energy trapping in the acceptor state, $|\text{acceptor}\rangle$, phenomenologically by introducing anti-Hermitian contributions in the Hamiltonian

$$\mathcal{H}_{\text{trap}} = -i\hbar\Gamma_{\text{trap}}/2|\text{acceptor}\rangle\langle\text{acceptor}|, \quad (3.3)$$

where Γ_{trap} defines the trapping rate. Similarly, we model radiative and non-radiative decays to the electronic ground state as exciton losses,

$$\mathcal{H}_{\text{loss}} = -i\hbar\Gamma_{\text{loss}}/2\sum_i|i\rangle\langle i|. \quad (3.4)$$

The rate $\Gamma_{\text{loss}}^{-1}$ defines the exciton lifetime. In this study we are interested in two different exciton propagation characteristics: the average transfer time, defined as

$$\langle t \rangle = \Gamma_{\text{trap}}/\eta \int_0^{t_{\text{max}}} dt t \langle \text{acceptor} | \boldsymbol{\rho}(t) | \text{acceptor} \rangle, \quad (3.5)$$

and the overall efficiency, defined as

$$\eta = \int_0^{t_{\text{max}}} dt \Gamma_{\text{trap}} \langle \text{acceptor} | \boldsymbol{\rho}(t) | \text{acceptor} \rangle, \quad (3.6)$$

which corresponds to the accumulated trapped population during the transfer process. For numerical evaluations, we replace the upper integration limit by t_{max} which is chosen such that the total population within the pigments at time t_{max} has dropped below 10^{-4} . The exciton dynamics is expressed in terms of the reduced density matrix, $\boldsymbol{\rho}(t)$, which can be computed with standard open quantum system approaches. Here, we use the HEOM approach which exactly accounts for the reorganization process,^{119,364–366} in which the vibrational

coordinates rearrange to their new equilibrium position upon electronic transition from the ground to the excited potential energy surface. The major drawback of the HEOM approach is the adverse computational scaling, which arises from the need to propagate a complete hierarchy of auxiliary matrices. We employ a high-performance implementation of HEOM integrated into the *QMaster* software package.^{337,359,367} A computationally much cheaper formalism, the Redfield approach, can be derived with the assumption of weak couplings between the system and the bath in combination with a Markov approximation.^{144,354} The secular approximation simplifies the equation even further and allows us to write the dynamics in the form of a Lindblad master equation. This strategy drastically reduces the computational demand compared to exciton propagation under the HEOM, which gives rise to the popularity of the secular Redfield equations. However, secular Redfield has been shown to underestimate the transfer times in certain light-harvesting complexes.³⁶⁸

3.3 DATA-DRIVEN APPROACH TO EXCITATION ENERGY TRANSFER

Several studies across various fields in recent years have demonstrated how ML models can be used to accelerate computations by several orders of magnitude at a reasonable level of accuracy. For example, formation free energies for catalyst surface chemistry were predicted with a Gaussian process (GP),³⁶⁹ kernel ridge regression methods were found to accurately predict atomization energies of small molecules,³⁷⁰ neural networks have been employed for the successful construction of various forms of transferable and non-transferable atomistic potentials,³⁷¹⁻³⁷³ atomic convolutional neural networks accurately reproduce protein-ligand binding affinities,³⁷⁴ and MLPs were trained to predict excited state energies in the context of exciton dynamics,³⁷⁵ as well as other electronic properties of small molecules.^{370,376} The study of EET typically involves two steps: first, an effective Hamiltonian describing the system parameters needs to be constructed, and second, transfer properties need to be computed from this effective Hamiltonian using open quantum system approaches (see Sec. 2.1). Recently, it has been demonstrated that ML can aid in the first step, the construction of the effective Hamiltonian, by predicting excited state energies of excitonic pigments from

Coulomb matrices.³⁷⁵

In the subsequent sections, we develop a ML framework based on MLPs to predict EET properties of excitonic systems from an effective Hamiltonian rather than obtaining them from computationally expensive quantum dynamics calculations. In future applications, this approach could facilitate large-scale screening, such as the search for best-performing devices or studies on structure-function relationships in natural light-harvesting. MLPs have been shown to generally perform well in supervised regression problems in chemistry.^{370,375} Further, we choose MLPs since there is no informative relation between neighboring elements in the Frenkel exciton Hamiltonian, which could be exploited by, *e.g.*, convolutional or recurrent neural networks, as excitonic sites can be numbered in arbitrary order. Overall, our procedure can be summarized as follows. We leverage standard open quantum system approaches to generate a database comprising of average transfer times and efficiencies for EET from a donor to a target pigment for a random set of Frenkel exciton Hamiltonians. The complete dataset is split into a training set, on which we train each MLP model, as well as a validation set for hyperparameter optimization and a test set to assess the predictive capabilities of the trained models. For training data selection, we will compare two strategies: (i) random selection of data points and (ii) selection of training data based on a principal component analysis (PCA), which allows us to identify and extract those data points covering the most information sampled in the dataset. As we show in Sec. 3.4, the latter strategy is of particular relevance if the space of transfer properties is not evenly sampled and many representatives in the training set contain redundant information. We use Bayesian optimization to identify the best architectures for our MLP models. The performance of each architecture is quantified by the average relative absolute error made when predicting transfer properties for the validation set. Finally, we run predictions on the test set to assess the ability of the optimized architectures to generalize to realizations that were neither employed for training nor validation. The source code for exciton transfer property predictions along with all trained MLP models, as well as the datasets generated in this study, are made available on GitHub.³⁷⁷

3.3.1 GENERATING THE EXCITATION ENERGY TRANSFER DATABASE

To demonstrate the capabilities of our ML approaches, we construct and study four datasets of randomly generated excitonic systems that are sampled around pigment-protein complexes found in natural light-harvesting. For our first dataset, we sample Hamiltonians around the FMO complex (Fig. 3.1), which frequently serves as the prototype light-harvesting complex (see Sec. 2.1). We construct three additional datasets that are motivated by the photosystem II of higher plants. For one set, we consider the eight pigments of the RC core, in which the primary step of charge separation is initiated through the electronically excited pigment Chl_{D1} .^{357,378} For the other two sets, the RC core is extended by including either light-harvesting complex CP47 or CP43 of photosystem II into the excitonic system. For simplicity, we refer to the dataset inspired by the CP43+RC (CP47+RC) complex as the CP43 (CP47) dataset from hereon. For each dataset, we generated 12,000 exciton Hamiltonians by uniformly sampling excited state energies and inter-site couplings from a fixed range of values, as is summarized in Tab. 3.1.

Table 3.1: Lower and upper limits in between which excited state energies ε and inter-site couplings V were sampled uniformly to generate the four datasets of this study. Each dataset consists of 12000 Hamiltonians with excited state energies and inter-site couplings within the reported ranges. Note, that the labels CP43 (CP47) denote datasets which are inspired by the CP43+RC (CP47+RC) biological complexes

Label	# Sites	ε_{low} [cm^{-1}]	$\varepsilon_{\text{high}}$ [cm^{-1}]	V_{range} [cm^{-1}]
RC	8	14800	15000	-50 ... 50
FMO	8	12000	12800	-100 ... 100
CP43	21	14800	15100	-60 ... 60
CP47	24	14500	15300	-100 ... 100

We compute exciton transfer times for all Hamiltonians in our datasets with the HEOM^{119,140,141} method, implemented in the *QMaster* software package, version 0.2.^{337,359,367} HEOM is a numerically exact method which accurately accounts for the reorganization process,^{118,364–366} during which the vibrational coordinates rearrange to their new equilibrium positions upon electronic transition from the ground to the excited potential energy surface. For all Hamiltonians we assumed identical Drude-Lorentz spectral densities, $J(\omega) = 2\lambda \frac{\omega v}{\omega^2 + v^2}$, to describe exciton-phonon interactions. More details on the Frenkel exciton Hamiltonian and the exci-

ton dynamics methods, as well as the definition of the transfer time and transfer efficiencies, are provided in the appendix of Ref. [330].

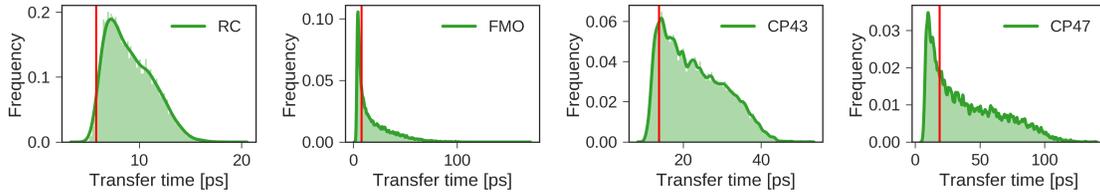


Figure 3.2: Distributions of exciton transfer times computed for all 12,000 generated exciton Hamiltonians for each dataset using the HEOM approach. Vertical red lines indicate the transfer time of the exciton Hamiltonian corresponding to the biological complex. In all calculations we use a trapping rate of $\Gamma_{\text{trap}}^{-1} = 1$ ps, an exciton life-times of $\Gamma_{\text{loss}}^{-1} = 0.25$ ns, and a temperature of $T = 300$ K. The parameters of the spectral density are set to $\lambda = 35 \text{ cm}^{-1}$, $\nu^{-1} = 50 \text{ fs}$. Reproduced from Ref. [330] with permission from the Royal Society of Chemistry.

Fig. 3.2 illustrates the transfer time distributions for all exciton Hamiltonians in each dataset. The transfer times for the Hamiltonians of the biological complexes are highlighted in every distribution. Excited states and inter-site couplings for the exciton Hamiltonians of the biological complexes are taken from literature.^{111,355–357} All population dynamics simulations are initialized as a fully populated site 1, serving as a donor, while site 3 acts as the acceptor which couples to an energy sink with trapping rate Γ_{trap} . Note that the labeling of the donor and acceptor state is without loss of generality as rows and columns of the Hamiltonian can be permuted accordingly, which effectively corresponds to a relabeling of the pigments. Since excited state energies and inter-site couplings are drawn from the same distributions for all sites in one dataset we did not explicitly account for the ordering ambiguity which arises, for instance, in the case of Coulomb matrices for which matrix entries depend on the particular types of atoms to which they correspond.³⁷⁰ We find large variations in the ranges of transfer times between the four datasets. The RC and CP43 datasets, both with relatively narrow ranges of excited state energies and site couplings, yield relatively short transfer times.

In contrast, we observe a broader spread in transfer times for the FMO dataset and the CP47 dataset, which is consistent with the broader range of excited state energies and site couplings that were sampled. The transfer times of the actual biological complexes lie

close to the mode of the distributions for all four datasets. This observation suggests that natural light-harvesting systems may not be selected explicitly for extraordinary transfer properties, as they exhibit transport characteristics that are just likely to occur, even for random perturbations of the exciton Hamiltonian. We note that providing a conclusive answer goes beyond the scope of the present manuscript, but could be the subject of a future, more detailed structure-function analysis. A recent evolutionary study for the FMO complex¹³¹ goes along a similar direction and suggests that the FMO complex has evolved towards stability to mutations rather than a selection of specific transfer characteristics.

3.3.2 PRINCIPAL COMPONENT ANALYSIS FOR IMPROVED TRAINING DATA SELECTION

We select the training sets for our MLP models following two methods for dataset splitting. In the simplest approach, we select the training set randomly from our created dataset. However, due to the nature of how we randomly sampled our Hamiltonians, the transfer characteristics are not distributed uniformly, and many representations of our Hamiltonians might be very similar and are thus expected to carry redundant information. As can be seen in Fig. 3.1, Hamiltonians yielding longer transfer time-scales are, for example, underrepresented in all four datasets. Our second approach follows a different path and implements a more sophisticated selection process. The idea is to add those Hamiltonians to our training set, which provide the most information. We perform a PCA on the 8,000 Hamiltonians containing dataset after separating 2,000 Hamiltonians each for validation and testing. We project each Hamiltonian onto a reduced space spanned by the most relevant principal components. The Hamiltonians for the training set are selected such that they are maximally separated in the reduced space. This procedure guarantees that our training set constitutes the most diverse samples.

3.3.3 SETUP OF THE MULTI-LAYER PERCEPTRON ARCHITECTURE

The architectures of our MLPs are designed for the supervised learning of EET properties. All exciton Hamiltonians were reshaped into vectors and provided as input features to the

MLPs, which were used to predict exciton transfer times and transfer efficiencies simultaneously. Since the input features of neural networks need to be of fixed size, we construct separate MLPs for each dataset to treat the different dimensionalities of the exciton Hamiltonians. Details on the rescaling of the input features and predicted output, as well as on the training procedure are provided in the appendix of Ref. [330]. The 12,000 Hamiltonians of each dataset were split into three sets: a training set of up to 8,000 Hamiltonians for training MLP model instances with particular hyperparameters, a validation set of 2,000 Hamiltonians used to evaluate the MLP architecture during optimization of the hyperparameters and a test set of 2,000 Hamiltonians to probe out-of-sample prediction accuracies. All constructed MLP models were trained with stochastic gradient descent with 200 data points per batch and the Adam optimizer,³⁷⁹ until the average relative absolute error (see Eq. 3.7) on the validation set increased over three full consecutive training epochs. Neuron saturation was avoided with L2 regularization on all weights of all neurons but the output neurons.

Table 3.2: Average relative absolute error $\Delta\tau$ (see Eq. 3.7) of exciton transfer times computed with HEOM and either, predicted by the trained neural networks (with/without PCA selection) or computed with secular Redfield. For all four datasets, we show the results of the training, validation, and test set separately. Smallest errors for each dataset are printed in bold

Dataset	Model	$\Delta\tau_{\text{train}}$ [%]	$\Delta\tau_{\text{valid}}$ [%]	$\Delta\tau_{\text{test}}$ [%]
FMO	Network (PCA)	4.53	4.38	7.41
	Network	10.53	10.75	11.56
	Redfield	9.70	9.96	9.60
RC	Network (PCA)	2.71	2.73	3.35
	Network	3.61	3.58	3.76
	Redfield	8.62	8.67	8.60
CP43	Network (PCA)	4.42	4.47	4.72
	Network	4.66	4.71	4.86
	Redfield	4.71	4.66	4.73
CP47	Network (PCA)	12.36	12.32	12.59
	Network	13.36	13.34	13.59
	Redfield	10.48	10.47	10.51

An essential component in developing accurate ML models consists in the choice of proper values for the model hyperparameters. For this MLP framework, we consider a total of six

hyperparameters, such as the initial learning rate, ν , for the Adam optimizer, and the regularization parameter, λ . We also included the number of MLP layers and the number of neurons per layer, as well as the activation functions for neurons in each layer, which were chosen from five different options. The only exception is the last layer, for which we always use the softplus activation function to constrain our MLP models to the prediction of always positive transfer times and efficiencies. Lastly, we treat the number of training points as a hyperparameter in order to study the effect of the variations in the number of training samples on the prediction accuracy. The set of hyperparameters to be optimized and their allowed ranges are summarized in the appendix of Ref. [330]. We employ a Bayesian optimization algorithm,³⁸⁰ to scan the space of hyperparameters for the most accurate model (see Sec. 2.3). This approach reduces the number of costly function evaluations under the assumption that the unknown function was sampled from a GP. In contrast to gradient or Hessian based optimization techniques, Bayesian optimization uses the information of all previously evaluated points and can thus find a good approximation to the minimum of non-convex functions in relatively few iterations. The model accuracy was defined as the average relative absolute error (see Eq. 3.7) in exciton transfer times predicted by the MLP and corresponding HEOM simulations for the validation set. All generated MLP models were constructed and trained with the same random seed. We carried out the Bayesian optimization of MLP hyperparameters in the SPEARMINT software package.²³⁷ MLP models were generated and trained using the Tensorflow package, version 1.0.³⁸¹

3.4 PREDICTION OF TRANSFER TIMES WITH NEURAL NETWORKS

In the subsequent discussion, we demonstrate the capabilities of our trained MLP models by analyzing the average relative absolute error,

$$\Delta\tau = \left\langle \frac{|t_{\text{HEOM}} - t_{\text{model}}|}{t_{\text{HEOM}}} \right\rangle_{\text{dataset}}, \quad (3.7)$$

between predicted exciton transfer times and the ones obtained with the numerically exact HEOM calculations. Although we restrict our discussion to transfer times, we note that similar conclusions hold for the analysis of the transfer efficiencies since both characteristics are strongly correlated. Tab. 3.2 summarizes the results for the predicted transfer times for our four generated datasets. The predictions are carried out with the Bayesian optimized MLP architectures, which show slight variations in their best-performing hyperparameters depending on the dataset at hand. However, for all datasets, the neural networks tend to prefer shallow but broad architectures comprising of only a few layers with each layer containing a larger number of neurons. More details on the procedure and results for the hyperparameter optimization can be found in the appendix of Ref. [330].

3.4.1 PREDICTION ACCURACIES OF TRAINED MULTI-LAYER PERCEPTRONS

Our trained MLP models predict exciton transfer times for out-of-sample Hamiltonians at almost the same accuracy as for Hamiltonians on which MLP parameters and hyperparameters were optimized (see Tab. 3.2). This observation demonstrates the ability of our MLP models to generalize to previously unseen data and to provide accurate out-of-sample predictions. Noteworthy, there is no significant asymmetry in the distribution of the relative absolute errors for the individual Hamiltonians or the training/validation and test set (see Fig. 3.3a). Therefore, the architectures of the neural networks are well-balanced and neither in the regime of over-fitting, which would result in a large discrepancy in errors between the training and validation sets nor did we over-optimize the neural network architecture during Bayesian optimization. Overall we find a high accuracy of our predictions and small average relative errors on the test sets which are in the range between 3.35 % for RC (PCA selected training set) and 13.59 % for the largest considered exciton system CP47 attached to RC (randomly selected training set). The CP47 dataset exhibits the most diverse transfer properties (see Fig. 3.2), which explains the larger average relative absolute errors in the predictions when compared to the other datasets. Prediction accuracies for exciton Hamiltonians with permuted rows and columns are reported in the appendix of Ref. [330]. We

find prediction accuracies similar to those achieved on the test sets for Hamiltonians with permutations not involving the source or target sites. The observed prediction errors are also consistent with the distance distributions of Frenkel exciton Hamiltonians for each of the four datasets, which indicates that MLP models generally benefit from a finer sampling of the input parameter space.

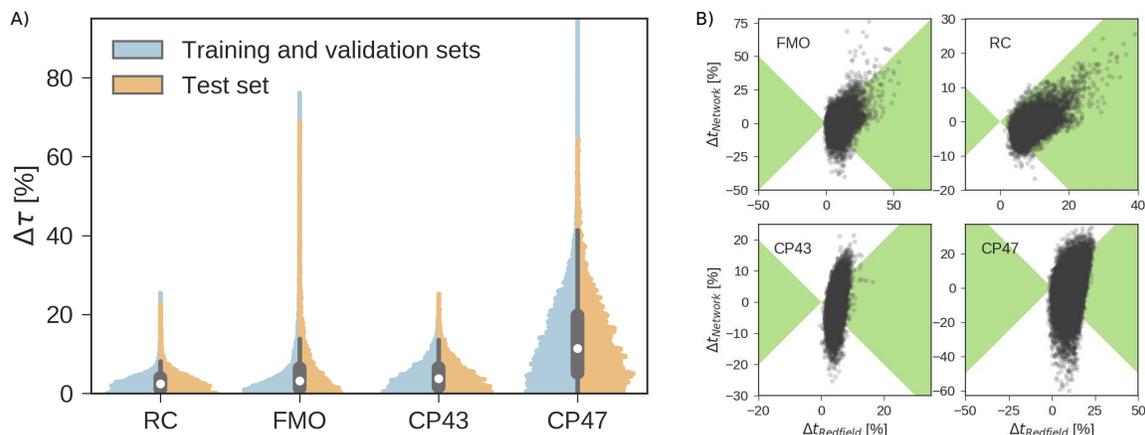


Figure 3.3: (A) Normalized distributions of the average relative absolute error of predicted exciton transfer times and exciton transfer times computed with HEOM. The left (blue) side of the plots illustrates the distributions of average relative absolute errors for predictions on the training and the validation set, while the right (orange) side of the plots illustrates the errors for predictions on the test set. (B) Relative errors in exciton transfer times computed with the hierarchical equations of motion (HEOM) approach and exciton transfer times computed with the secular Redfield approach and predicted by neural networks, respectively. Displayed are relative deviations for all four datasets: the Fenna-Matthews-Olson (FMO) complex, the reaction center (RC) core, the RC with the CP43 complex, and the RC with the CP47 complex. Regions in which the absolute of deviations of neural network predicted transfer times from HEOM computed transfer times are shorter than deviations for Redfield are shaded in green. Reproduced from Ref. [330] with permission from the Royal Society of Chemistry.

The accuracy of the predictions can be enhanced with the more sophisticated PCA selection of the training set without the need to generate additional computationally expensive data points. The level of improvement of the PCA selection over a random selection of the training set differs for the four complexes. In general, we find that MLPs can be trained to almost equal accuracy with either selection method. The highest benefit of the PCA selected training set is observed for the FMO, and CP47 dataset, which are not only the most diverse ones out of our four datasets but are biased towards Hamiltonians showing fast transfer. As intuitively expected, selecting training points based on PCA is most advantageous for datasets with relatively unevenly sampled feature spaces.

3.4.2 COMPARING DATA-DRIVEN PREDICTIONS TO SECULAR REDFIELD RESULTS

Next, we provide a context for the observed MLP prediction accuracies by comparing them to the errors made by the frequently employed secular Redfield method, which is derived from second-order perturbation theory in the system-bath interaction in combination with a Markov approximation. Accuracies of the transfer times for both, the secular Redfield calculations and the MLP predictions are evaluated according to Eq. 3.7. Here, the HEOM calculations again serve as ground truth. For the datasets inspired by the smaller exciton systems FMO and RC, the trained MLPs outperform secular Redfield, even for out-of-sample predictions, whereas for the datasets around larger systems both approaches are similarly accurate. For example, in the case of the biological exciton Hamiltonian of the FMO complex, HEOM reveals a transfer time of 7.95 ps. The trained MLP model predicts a transfer time of 7.52 ps which is slightly more accurate than secular Redfield calculations that result in 7.48 ps. Exciton transfer times obtained for all four biological complexes with all three approaches are reported in the appendix of Ref. [330]. However, while the MLP prediction takes about 5 ms, secular Redfield calculations took about 14.5 min on a single CPU. We conclude that our trained MLP predictions are competitive to secular Redfield calculations in terms of their accuracy, but (once trained) come at a significantly reduced computational cost.

Besides analyzing the accuracy in terms of averaging over all realizations in the datasets, we compare the relative errors in transfer time for secular Redfield and the MLP predictions in more detail on the level of individual Hamiltonians. Fig. 3.3b depicts scatter plots where the horizontal axes measure the accuracy of secular Redfield calculations and the vertical axes reflect the accuracy of MLP predictions for MLPs trained on the PCA selected datasets. We do not distinguish between training, validation, and test set and show the complete dataset. Almost all the Hamiltonians show a $\Delta t_{\text{Redfield}} = (t_{\text{HEOM}} - t_{\text{Redfield}})/t_{\text{HEOM}} > 0$, which demonstrates that secular Redfield systematically underestimates transfer time scales. On the other hand, the predictions under- as well as overestimate transfer time-scales yielding

a more symmetrical distribution along the horizontal axis. For the RC (FMO) dataset, more than 95 % (80 %) of the Hamiltonians fall into regions marked as green, for which the neural networks provide higher accuracy than secular Redfield. For all other datasets, secular Redfield and the MLP predictions are equally likely to give better results, with about 59 % (57 %) of the Hamiltonians for CP43 (CP47) falling within the green shaded region. This observation is in agreement with our average relative absolute errors listed in Tab. 3.2. We did not observe any cases for which the MLPs show relative errors that significantly exceeded any of the secular Redfield ones.

3.5 CONCLUSION

In this study, we have outlined how ML approaches can be used to bypass computationally costly simulations of open quantum system dynamics in the context of EET. Overall we find that MLPs are capable of predicting transfer times for excitonic systems at higher or comparable accuracy than the frequently used secular Redfield approach, albeit at much lower computational costs. We conclude that MLP models are a promising alternative for extracting excitation energy transfer properties when compared to frequently used rate equation methods. The presented approach is of particular interest for large-scale analyses of the structure-transport relationship in excitonic systems. An area of great interest in excitonics is the study of the dynamics of charge dissociation at the interface present in bulk heterojunction photovoltaics.^{382,383} We believe a tool like this will help in the rapid screening of material properties in the mesoscale and therefore streamline the search for high-performance organic photovoltaic (OPV) systems.³⁸⁴ Once trained, evaluations of MLP models come at almost no additional cost. Our four generated MLP architectures (each optimized for one of the four datasets) predict transfer times for an aggregated set of 48,000 exciton Hamiltonians just within a few seconds, while the corresponding quantum dynamics simulations take several GPU (CPU) years for the HEOM (secular Redfield) calculations. Our trained MLP models extend well to out-of-sample predictions for exciton Hamiltonians that are close to the sampled parameter regime. Nevertheless, to employ MLPs on parameter

regimes beyond those probed in the existing database requires running computationally expensive exciton dynamics for a few thousand Hamiltonians in order to extend our training set. To avoid this bottleneck, a potential strategy could be to leverage already existing data, *e.g.*, produced by a user community of existing software packages such as *QMaster*. However, such data can be quite diverse. To this end, future research needs to focus on more general neural network architectures that accurately predict transfer times for flexible spectral density parameters as well as differently sized exciton systems.

4

From absorption spectra to charge transfer in oligomeric nanoaggregates with machine learning

Apart from minor modifications, this chapter originally appeared as:³⁸⁵

From absorption spectra to charge transfer in nanoaggregates of oligomers with machine learning. Loïc M. Roch, Semion K. Saikin, Florian Häse, Pascal Friederich, Randall H. Goldsmith, Salvador León and Alán Aspuru-Guzik. *ACS Nano*. in press, 10.1021/acsnano.0c00384 (2020).

Reproduced with permission from Ref. [385] Copyright 2018 American Chemical Society

ABSTRACT

The fast and inexpensive characterization of materials properties is a crucial element to discover novel functional materials. In this work, we suggest a data-driven approach with three classes of Bayesian machine learning models to correlate electronic absorption spectra of nanoaggregates with the strength of intermolecular electronic couplings in organic conducting and semiconducting materials. As a specific model system, we consider poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS), a cornerstone material for organic electronic applications. Specifically, we study the couplings between charged dimers of closely packed poly(3,4-ethylenedioxythiophene (PEDOT) oligomers that are at the heart of the material's unrivaled conductivity. We demonstrate that machine learning approaches can identify correlations between coupling strengths and optical absorption spectra. We also show that data-driven models can be trained to be transferable across a broad range of spectral resolutions and that the electronic couplings can be predicted from the simulated

spectra with an 88 % accuracy when used as classifiers. Although the machine learning models employed in this study were trained on data generated by a multi-scale computational workflow, their demonstrated robustness suggests that they generalize well to experimental data. This study illustrates that accessible properties of a system can be used as a proxy to determine more inaccessible properties using data-driven approaches to construct empirical relations between these two properties. As such, data-driven tools can enable new routes to experimentation and materials discovery.

4.1 INDIRECT PROPERTY CHARACTERIZATIONS FOR ORGANIC ELECTRONICS

Organic materials are attractive for optoelectronic device applications, notably due to their low fabrication cost and their relative ease to produce and characterize.³⁸⁶ Not only can the structural properties of these materials be tuned through the functionalization of molecules,³⁸⁷ but they are also composed of elements which are earth-abundant. In contrast to conventional inorganic electronic materials, organic compounds provide flexibility,³⁸⁸ biocompatibility³⁸⁹ and biodegradability,³⁹⁰ as well as self-healing properties (see Sec. 2.1).^{391,392} Organic conducting and semiconducting materials hold promises for several application niches, including next-generation wearable and printed photovoltaics,^{393,394} fuel cells,^{395,396} thermoelectrics,³⁹⁷⁻⁴⁰⁰ and others,^{401,402} which demonstrates their critical importance to solving immediate societal challenges (see Chapter 1).

One of the fundamental challenges to the design of organic optoelectronics lies in the intrinsic structural disorder of these materials. This disorder emerges on multiple length scales starting from the conformations of single molecules and the nearest-neighbor packing to the formation of multi-molecule domains and nanocrystals. The electronic properties of organic materials are highly sensitive to the packing of composing molecules, hence dependent on the processing conditions.⁴⁰³ Fast optical probing of local electronic couplings can benefit both applied and fundamental research. On the one hand, such a method brings the possibility to combine continuous testing of devices with roll-to-roll device manufacturing technologies.⁷ On the other hand, optical characterization techniques can advance our

understanding of charge transport in organic structures. In particular, ultraviolet-visible spectroscopy (UV/Vis), X-ray photoelectron spectroscopy (XPS), and Raman scattering measurements of thin films of conductive polymers can provide insights to composition and electronic structure, including the nature of charge carriers (see Chapter 11).^{404,405}

The microscopic structure of molecule and polymer packing is challenging to measure directly. Obtaining optical spectra, however, such as infrared spectroscopy (IR) absorption, Raman scattering, electronic absorption, and fluorescence is more straightforward and requires substantially less experimental effort. Both electronic and optical properties are influenced by the microscopic molecular packing. In the most straightforward qualitative picture, the proximity of two molecules yields an overlap of electronic clouds, which results in charge transfer. This proximity also leads to a Förster coupling between electronic excitations, which can be observed as changes in the lines in the electronic absorption spectra.⁴⁰⁶ Moreover, weak charge-transfer excitations can be developed with sufficient electronic coupling between molecules. Because both effects are caused by molecular interactions, in principle, it is possible to find a data-driven model that correlates them to enable indirect characterizations of electronic properties *via* optical measurements. The conventional computational approach involves three steps: (i) building physical models that describe both properties of interests; (ii) fitting the parameters of the models to experimental data, *e.g.*, absorption spectra; and (iii) using the fitted models to describe the other property, *e.g.*, conductivity. Such an approach might be challenging since the relations between these properties can be too complex to derive a tractable or straightforward physical model and the optical characterization of the material might not encode all relevant electronic information.

Herein, we report an alternative approach, where data-driven models substitute the aforementioned physical model. To this end, we design a multi-scale computational workflow where the first three steps – force-field calculations, molecular dynamics (MD) simulations, and quantum-based approaches – generate empirical evidence. In this work we employed machine learning (ML) algorithms to identify correlations between the two properties of interest, *i.e.*, the strength of intermolecular coupling and electronic absorption spectra. Then,

we used the trained ML models as relative classifiers of the coupling strength of a given spectrum with respect to a reference coupling, which is to be defined by a scientist for the application at hand. As a model system, demonstrating the reliability of the classifier to identify structures with strong electronic couplings from their absorption spectra, we study pairs of PEDOT oligomers. PEDOT is one of the most technologically-developed conducting polymers. Owing to its high hole conductivity and optical transparency in a doped state⁴⁰⁷ it is widely used for transparent contacts in photovoltaic devices, touch screens, and light-emitting diodes.⁴⁰⁸ PEDOT-based materials are frequently fabricated as a mixture with polystyrene sulfonate (PSS) polymers, PEDOT:PSS. In this mixture, PEDOT oligomers transfer the charges while PSS chains play the role of a solid electrolyte. This material becomes conductive at high concentrations of dopant.⁴⁰⁹

Although multiple experimental studies have addressed the molecular organization of PEDOT-based materials,^{404,405,407,410–416} the microscopic electronic states that lead to high conductance and the interplay between these states, optical properties, and the material structure have yet to be determined. The critical factor for practical applications of PEDOT-based materials lies in understanding the relations between their solid-state packing and their unique electronic properties. The main obstacle to elucidating this relationship is the strong structural disorder that appears on multiple length scales and is highly sensitive to the thin film preparation procedure.⁴⁰⁷

Hereafter, we demonstrate that our data-driven models confirm the existence of correlations between the coupling strengths and the electronic absorption spectra. We also show the robustness due to the Bayesian formulation of our ML models (see Sec. 2.2) with respect to potential spurious statistical correlations to capture the relevant physical relations. Finally, we use the ML models as classifiers to determine whether a given absorption spectrum relates to a coupling strength above or below an *a priori* selected reference coupling strength. Such an approach has proven to be reliable and robust, reaching an average error rate of only 12% when employing a Bayesian convolutional neural network (CNN). The importance of such a classifier becomes apparent in the context of the self-driving laboratories,^{7,81,417,418}

where the goal of the experimentation process is embodied as an optimization procedure (see Part III) to identify fabrication procedures for aggregates yielding high conductivities.

4.2 COMPUTATIONAL APPROACHES TO CHARGE TRANSFER STUDIES

This section details the computational workflow designed to generate the data and to correlate electrical and optical properties using ML approaches. The workflow is depicted in Fig. 4.1. Each of the five composing steps (*i.e.*, initial structures, refinement, MD simulation, physical models, ML models in Fig. 4.1a) are described in their corresponding subsections. Hereafter, we assume that the packing of conjugated oligomers in the solid PEDOT:PSS mixture depends on the initial preparation procedure and post-processing steps. For PEDOT:PSS, such steps are critical to achieving peak performance. It is hypothesized that solid PEDOT:PSS films consist of grains with a hydrophobic and highly conductive PEDOT-rich core and a hydrophilic insulating PSS-rich shell.^{411,419} This phase segregation of PEDOT and PSS occurs on a scale beyond current computational capabilities and, thus, is not captured by our model. Nonetheless, our computational workflow allows us to study the disorder within each of these grains.

INITIAL STRUCTURES. To screen the orientation stability of the PEDOT:PSS complex within the grains, 100 starting structures were generated with PACKMOL.⁴²⁰ Each of the complexes consisted of one PEDOT chain with eight 3,4-ethylenedioxythiophene units ($n = 8$, Fig. 4.1b) carrying two positive charges, and two PSS chains consisting of three PSS units ($m = 3$, Fig. 4.1b) with one negative SO_3^- and two SO_3H groups per chain. In the generation of the initial structures, we constrained the SO_3^- group of the PSS chain to point towards the positive PEDOT chain. These initial structures were then optimized using a classical force field (FF) approach.

REFINEMENT. The 20 energetically most stable PEDOT:PSS complexes obtained from the initial structure search were relaxed at the B97-D/6-31G(d,p) level of theory in the

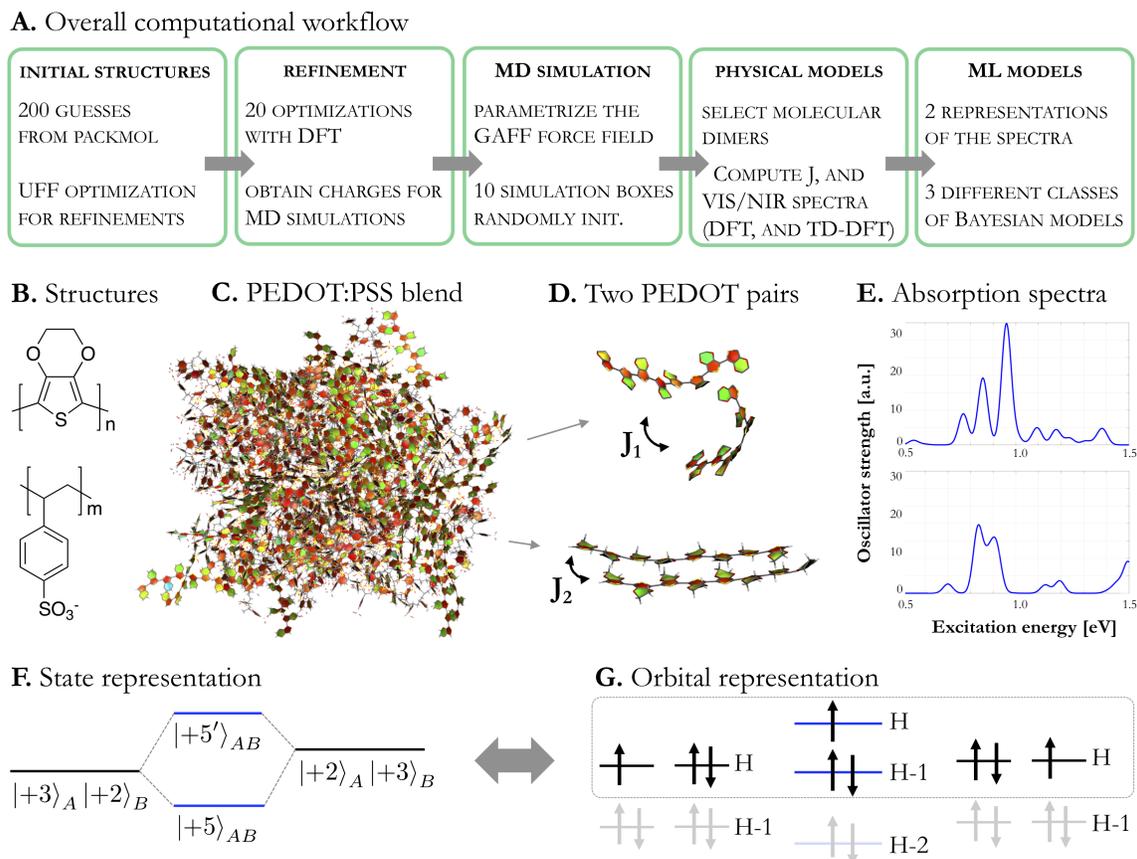


Figure 4.1: The computational pipeline used, from structure generation to correlating the electronic coupling, J , and electronic absorption spectra. (A) General workflow highlighting the steps and summarizing the methods involved. (B) Structures of PEDOT and PSS, represented in the top and bottom panels, respectively. (C) One of the ten supercells of PEDOT:PSS blends. (D) Two distinct example pairs of PEDOT oligomers extracted from the PEDOT:PSS bulk. J_1 and J_2 denote the coupling strengths for each of the pairs. (E) Associated simulated electronic absorption spectra. (F) State, and (G) orbital representations of the monomers and dimers involved in the calculation of the coupling strength, J . Reproduced from Ref. [385] with permission from the American Chemical Society.

gas phase, using the Gaussian software package.⁴²¹ Note that the influence of the solvent was found to be negligible and that the performance of the B97-D functional on geometries has already been assessed in previous work.⁴²² Single point energy calculations at the HF/6-31G(d,p)//B97-D/6-31G(d) level were performed on the 20 relaxed complexes to parameterize the charges for the MD simulations, as customary with the generalized Amber force field (GAFF).⁴²³

MD SIMULATIONS. Ten MD simulations were carried out on PEDOT:PSS model systems in periodic cubic boxes of size* $61.32 \text{ \AA} \times 61.32 \text{ \AA} \times 61.32 \text{ \AA}$ for a density of *ca.* 1.4 g/cm^3 , with the LAMMPS software package.⁴²⁴ The GAFF was chosen to describe the systems. PEDOT oligomers with eight repeat units ($n = 8$, Fig. 4.1b) and a +2 charge were considered, while the PSS atactic chains consisted of 20 repeat units ($m = 20$, Fig. 4.1b), with four deprotonated units randomly distributed in the sequence of each chain. Note that the PSS chain length was increased from $m = 3$ to $m = 20$ to better represent experimental blends. Additional details can be found in the appendix of Ref. [385].

PHYSICAL MODELS TO COMPUTE J , AND SIMULATE THE ELECTRONIC ABSORPTION SPECTRA. For two interacting PEDOT monomers, denoted A and B from hereon, the strength of the charge transfer integral, J , can be determined in several ways.^{425–434} Bi-polaron charge transport models have previously been discussed in the case of PEDOT systems.⁴³⁵ For the sake of simplicity, we assume a single-polaron transport model. Nonetheless, the ML models used for the prediction of electronic couplings are agnostic to the type of transport and will learn correlations between J and electronic absorption spectra independently of the type of the charge transport model. To calculate the coupling J , we used the framework of a tight-binding formalism^{436,437} as well as a Kohn-Sham orbital based method.^{426,432} In both cases, orbital energies were obtained at the B3LYP/def2-SV(P) level of theory. The choice of basis set and functional balances computational cost and accuracy. Detailed results on the performance of the def2-SV(P) results can be found in the appendix of Ref. [385].

The nearest-neighbor PEDOT dimers were extracted across the ten simulation boxes. Any pair of PEDOT molecules having at least two heavy (*i.e.*, non-hydrogen) atoms at a distance closer than 4 \AA is selected. The cutoff distance was taken to be comparable to the sum of the van der Waals radii of these heavy atoms. This procedure lead to the selection of 1,420 PEDOT pairs, ulteriorly used to model the strength of the charge transfer, and to simulate the electronic absorption spectra.

* Note: average size of the different boxes for the duration of the constant-pressure simulations

The tight-binding formalism is based on the change of orbital energies when going from isolated monomers to dimer systems,⁴³⁸

$$J = \sqrt{\left(\Delta\epsilon_{|AB\rangle,+5}^{H,L}\right)^2 - \frac{1}{4} \left[\left(\epsilon_{|A\rangle,+3}^{H,H-1} + \epsilon_{|B\rangle,+2}^{H,H-1}\right) - \left(\epsilon_{|A\rangle,+2}^{H,H-1} + \epsilon_{|B\rangle,+3}^{H,H-1}\right) \right]^2}, \quad (4.1)$$

where $\Delta\epsilon_{|AB\rangle,+5}^{H,L}$ is the splitting between the HOMO and the HOMO-1 level of the dimer AB with the charge $q = +5$, and $\epsilon_{|i\rangle,+q}^H$ denotes the HOMO energy of monomer i , carrying charge $+q$. Note that the correction due to the offset between the HOMOs of the monomers is negligible; hence, J is mostly governed by the splitting $\Delta\epsilon_{|AB\rangle,+5}^{H,H-1}$. This formalism considers frontier orbitals assuming that only the highest occupied orbitals are hybridized due to the electronic coupling between the oligomers (see Fig. 4.1f,g). Both the presence of non-equilibrated charges and non-zero spin increases the complexity of the model. Nonetheless, our interpretation of Eq. 4.1 is a lower estimate for the electronic coupling between the oligomers. Note that this model can also be used to describe bi-polaron transport.

The second approach to calculating charge transfer integrals uses the Kohn-Sham orbitals of isolated monomers as well as the Fock matrix and the overlap matrix of the dimer systems,^{426,432}

$$J_{AB} = \frac{F_{AB} - \frac{1}{2}(F_{AA} + F_{BB})S_{AB}}{1 - S_{AB}^2}. \quad (4.2)$$

The matrix elements $F_{AB} = \langle A|F_{\text{dimer}}|B\rangle$ and $S_{AB} = \langle A|S_{\text{dimer}}|B\rangle$ are calculated using the Fock and overlap matrices of a molecular dimer system with a charge of $+4$. The states $|A\rangle$ and $|B\rangle$ are the highest occupied molecular orbitals of the doubly positively charged monomers. The electronic absorption spectra of the 1,420 PEDOT pairs were simulated using the computed TD-CAM-B3LYP/def2-SV(P) transitions in the gas phase. Note that for the simulation of the spectra, the influence of solvation was found to be negligible. We employed a Lorentzian broadening to the time dependent density functional theory (TD-DFT) transitions. This translates the discrete oscillator strengths, f , and transition energies, ω , to continuous spectra to resemble experimental outcomes. The broadening was chosen to be 50 meV.⁴³⁹

Some insights about correlations between electronic absorption spectra and intermolecular charge transfer can be obtained only for the case when the coupling is weak. A completely relaxed doubly charged oligomer composed of eight to ten units would have a strong electronic transition at about 0.9 – 1.0 eV.⁴¹⁹ This transition is predominately composed of HOMO and LUMO orbitals. For an oligomer with an odd number of charges, additional HOMO-1 \rightarrow HOMO transitions appear at lower frequencies of *ca.* 0.5 eV. The low-frequency part of the spectra of dimers without the interaction should thus be composed of three lines – a low-frequency, weak transition, and a strong doublet. Electronic and excitonic interactions between the molecules further modulate the spectra. Specifically, weak charge transfer transitions should appear at the low-frequency tail of the spectra. The intensities of these transitions should be sensitive to the coupling strength, while their frequencies should be more stable as determined by the alignment of the molecular energy levels. However, this intuitive picture fails for intermediate and strong intermolecular couplings. In the latter case, the intramolecular states hybridize, which in turn leads to a realignment of the energy levels and a redistribution of the oscillator strengths among multiple transitions. An advantage of our ML approach is that it allows capturing correlations between the electronic interaction and optical spectra independently of the coupling regime.

ML MODELS TO IDENTIFY CORRELATIONS BETWEEN ELECTRONIC ABSORPTION SPECTRA AND COUPLING STRENGTHS. Correlations between electronic absorption spectra and coupling strengths were identified with ML models at different levels of complexity. To mimic experimental conditions, we encoded the Lorentzian broadened absorption spectra based on their intensities at specific frequencies, using a 1 meV binning on the considered frequency domain (0.5 eV to 1.5 eV). A total of 170 (11.97%) of the 1,420 spectra-coupling pairs were randomly selected to construct a test set. The remaining 1,250 (88.03%) of the dataset were used for ten-fold cross-validation. The size of the dataset motivates the use of Bayesian models for the robust and transferable identification of relevant physical correlations.

Specifically, we employed three different classes of Bayesian models (see Fig. 4.2): (i)

Bayesian linear regression models assume a linear dependence of coupling strengths on electronic spectra and thus present the simplest approximation; (ii) Bayesian multi-layer perceptrons (MLPs) are Bayesian generalizations of conventional deterministic MLPs with similar flexibility to model non-linear relations while retaining the robustness to overfitting of Bayesian methods; and (iii) Bayesian one-dimensional CNNs are special cases of MLPs which have the potential to efficiently exploit spatial correlations in the presented features due to their sparse nature. All models were set up to predict coupling strengths directly from the intensities of the associated absorption spectra at different frequencies. Additionally, we constructed Bayesian MLPs, which are trained on a compressed representation of the electronic absorption spectra obtained from principal component analysis (PCA). All models were trained based on an early-stopping criterion. Hyperparameters for all four models are optimized in a random grid search, and the details are provided in the appendix of Ref. [385].

ML MODELS AS RELATIVE CLASSIFIERS.

The aforementioned ML models trained for predicting absolute values of coupling strengths from electronic absorption spectra can be used to classify the conductivity associated with the electronic absorption spectra of the materials. Instead of asking for the absolute value of the coupling strength, the model provides an estimate for whether the considered coupling is above or below a reference. This reference is a hyperparameter defined by a scientist as a threshold for high and low values of J .

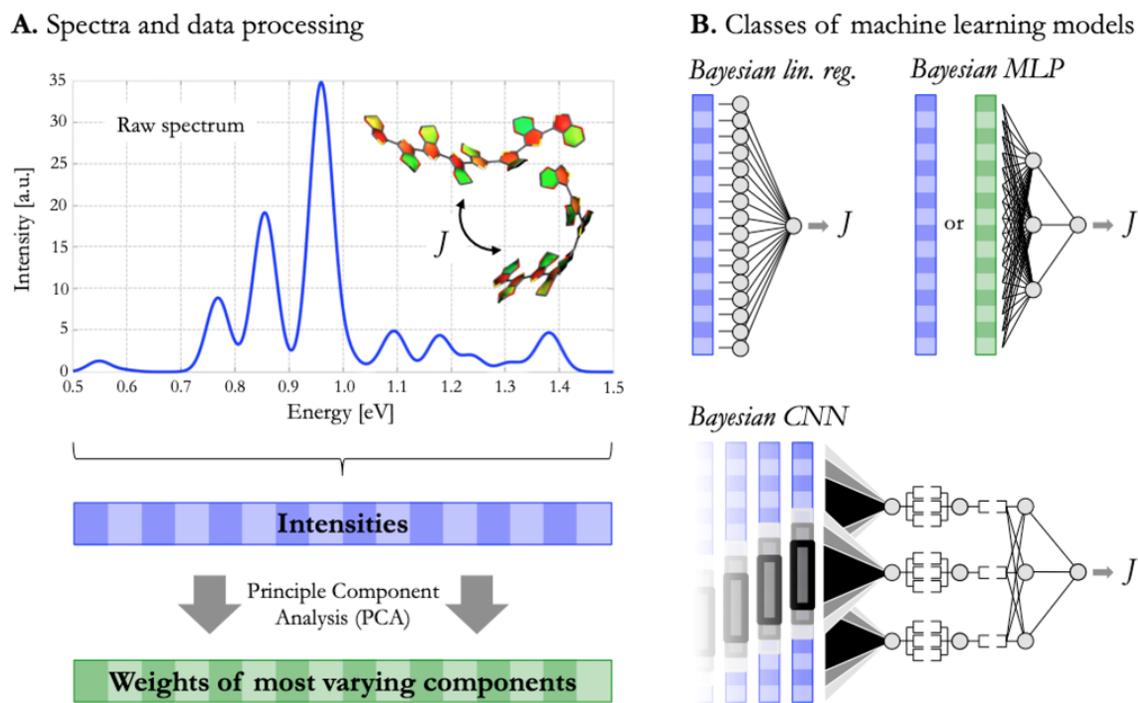


Figure 4.2: Schematic representation of (A) the data processing and (B) the three classes of ML models, depicting Bayesian linear regression models, Bayesian multi-layer perceptrons (MLPs), and Bayesian one-dimensional convolutional neural networks (CNN). Note that the MLPs are trained either on the raw intensities (blue array) or on the compressed representation after a principle component analysis (green array). Reproduced from Ref. [385] with permission from the American Chemical Society.

4.3 RESULTS AND DISCUSSION

We begin by discussing the performance of the ML models to identify correlations and to predict absolute values of coupling strengths from electronic absorption spectra. Fig. 4.3 illustrates the accuracies of all four models to predict coupling strengths computed from the tight-binding formalism with all four models after full hyperparameter optimizations. We proceed by detailing the tests designed to assess the performance of the ML models to capture relevant physical correlations. We continue our discussion with the results obtained when the ML models are used as relative classifiers, where associated error rates are reported in Fig. 4.4. We also highlight the practicality of such an approach in autonomous discovery applications with self-driving laboratories. Finally, we discuss the robustness of our ML models upon variations of the peak broadening. Fig. 4.5 illustrates the performance of our ML models at different broadenings.

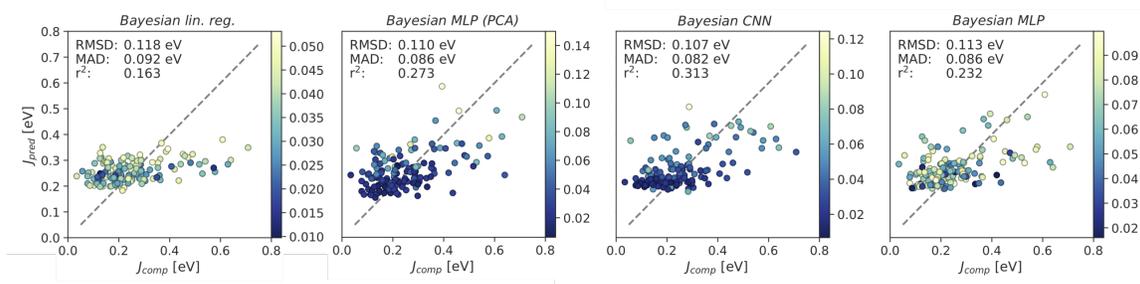


Figure 4.3: Coupling strengths predicted by the employed ML models in comparison to coupling strengths obtained with the tight-binding formalism. All reported predictions are shown for the test set, and were obtained from averaging 200 prediction samples from each of the models. Prediction uncertainties are color-coded. We report three comparative metrics to assess the prediction accuracies of each model: root mean square deviation (RMSD), mean absolute deviation (MAD), and coefficients of determination (r^2). Based on all three metrics, Bayesian CNNs provide the most accurate predictions. Reproduced from Ref. [385] with permission from the American Chemical Society.

4.3.1 CORRELATIONS BETWEEN ELECTRONIC SPECTRA AND COUPLING STRENGTHS

All constructed ML models predict coupling strengths at positive coefficients of determination (r^2), which indicates that the coupling strengths indeed correlate with the electronic absorption spectra of the PEDOT dimers and that the models are capable of identifying this

correlation. Bayesian CNNs provide the most accurate predictions based on all computed comparative metrics ($r^2 = 0.313$, RMSD = 0.107 eV, MAD = 0.082 eV) and Bayesian linear regression yields the least accurate predictions ($r^2 = 0.163$, RMSD = 0.118 eV, MAD = 0.092 eV). We further observe an improved prediction accuracy when compressing the electronic absorption spectra *via* PCA for the Bayesian MLP models.

Despite the relatively small size of the dataset, we found that the studied ML models, most notably the Bayesian CNN, present an efficient approach to identifying the physically relevant correlations. Estimates of the sampling efficiency of the Bayesian CNN model suggest that it can be trained to reach similar prediction accuracies with only 850 instead of 1,250 training points. No significant improvement in the prediction accuracy is observed when increasing the size of the training set beyond 850 examples. This observation, in conjunction with the generalization of the models observed for the test set predictions, indicates that the Bayesian CNN exploits all identifiable correlations to their full extent.

To ensure that our ML models did not capture spurious correlations that could arise from the methods and formalisms employed to compute the electronic absorption spectra and model the coupling strengths, we tested for the nature of the identified correlations by training the Bayesian CNNs to predict couplings strengths obtained with the Kohn-Sham orbital formalism from the same electronic absorption spectra. The trained Bayesian CNNs achieve prediction accuracies of $r^2 = 0.264$ on the same test set. We also constructed a hybrid dataset, where half of the couplings are randomly chosen from the tight-binding formalism and the other half from the Kohn-Sham orbital formalism. Again, the trained Bayesian CNNs achieve prediction accuracies of $r^2 = 0.280$ indicating that the presented ML models do not capture potential spurious statistical correlations but extract the relevant physical correlations.

4.3.2 MACHINE LEARNING MODELS AS RELATIVE CLASSIFIERS

The practicality of the presented ML models becomes apparent when weakening the requirement of accurate absolute predictions of the coupling strengths to accurate relative

predictions, which are of interest in discovery applications. Rather than requesting an estimate for the precise numeric value of the coupling strength, the trained model is used to determine if a given absorption spectrum relates to a coupling strength above or below an *a priori* selected reference. For such scenarios, the prediction accuracy of the model can be assessed by treating it as a binary classifier to determine true positive and true negative rates for different reference coupling strengths. We assess the prediction accuracies of such relative classifiers by estimating the probability of the model to make a correct prediction, *i.e.*, predicting the coupling to be above the reference when it is above or predicting the coupling to be below the reference when it is below (see Fig. 4.4). These probabilities are estimated for different reference coupling strengths, spanning the entire range of coupling strengths computed with the tight-binding formalism.

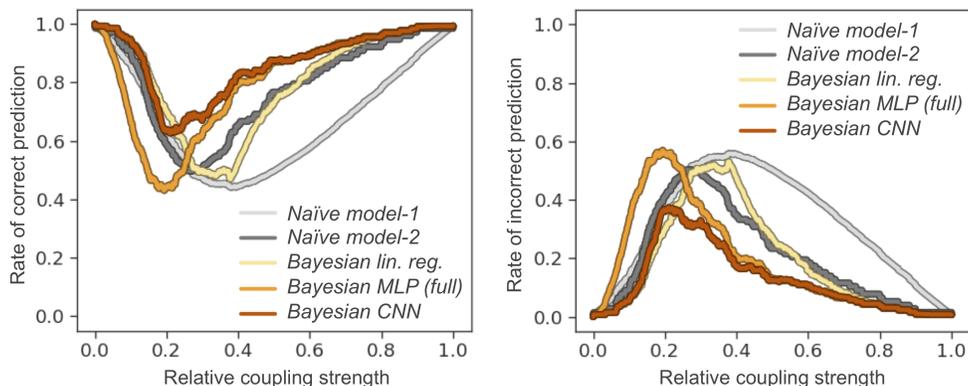


Figure 4.4: Probabilities to make correct predictions for J regarding the order with respect to another reference J (left panel) vs. making incorrect predictions (right panel). Reproduced from Ref. [385] with permission from the American Chemical Society.

Fig. 4.4 illustrates the error rates, *i.e.*, wrongly predicting a coupling strength to be above or below the considered reference coupling strength, for the trained ML models along with two naïve models for comparison. The most naïve model draws random samples from a uniform distribution to predict the coupling strength for a given electronic absorption spectrum (model-1, depicted in light grey in Fig. 4.4). A slightly more sophisticated naïve model predicts by drawing random samples from the distribution of coupling strengths (model-2, depicted in dark grey in Fig. 4.4). We find that the error rates for the two naïve models

yield the most substantial error rates: 31.9% and 20.2%. Bayesian linear regression scored an average error rate of 19.0%, which, despite its simplicity, already provides an advantage over naïve models and captures some of the relevant correlations in the dataset. The lowest error rate is observed for the Bayesian CNN with a 12.6% average error for coupling strengths chosen within the range of smallest and largest computed coupling strengths. Additionally, it is noted that the Bayesian CNNs never exceeds an error of 38% for any chosen reference coupling. If the focus of the discovery process is to identify fabrication procedures and post-processing steps leading to large coupling strengths (above 0.6 eV), the Bayesian CNN yields error rates of less than 10%. We suggest that the trained ML models can be applied to classify the coupling strengths with respect to a reference coupling strength with reasonable confidence.

4.3.3 ROBUSTNESS OF THE MACHINE LEARNING MODELS

Finally, we estimate the dependence of the model performances on the particular choice of the peak broadening. While we demonstrated that for one particular choice, the trained ML models are indeed capable of identifying the relevant correlations between the electronic absorption spectra and the coupling strengths, experimentally obtained electronic absorption spectra might be noisy and feature peaks at slightly varying broadenings. The robustness of the model predictions for different broadenings is tested by predicting coupling strengths from electronic absorption spectra at different broadenings, ranging from 5 meV to 1,000 meV. Note that while a 10 meV broadening is too small for a room temperature measurement, a 200 meV broadening would correspond to an unphysically fast dephasing rate. Fig. 4.5 summarizes the coefficients of determination for predictions of coupling strengths from electronic absorption spectra generated at different broadenings.

We find that the prediction accuracies of the Bayesian linear regression model are mostly insensitive to the particular broadening value (yellow trace). We only observe degradations in the predictive power for very small broadenings below 10 meV and very large broadenings above 200 meV. Bayesian MLPs (turquoise trace) show faster degradations in their predic-

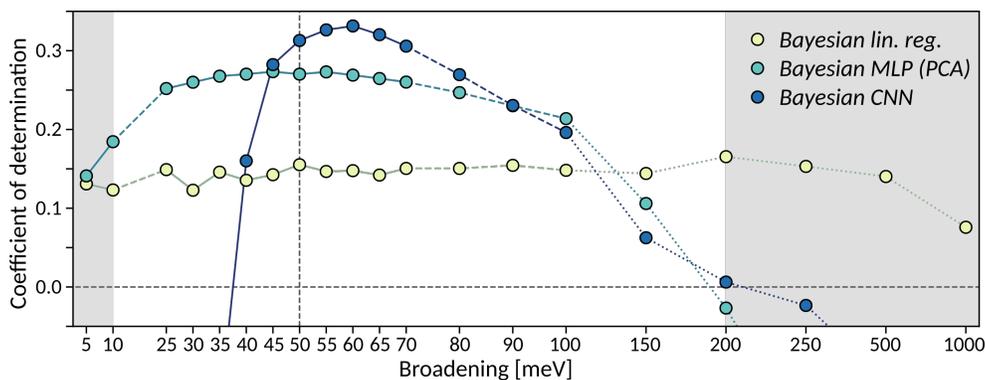


Figure 4.5: Prediction accuracies measured with r^2 coefficients for all models when predicting couplings from spectra at different broadenings. All three models depicted have been trained on a spectra at a 50 meV broadening (dashed vertical line). The regime ranging from 10 meV to 200 meV (highlighted in white) corresponds to the expected experimental resolution: while a 10 meV broadening is too small for a room temperature measurement, a 200 meV broadening corresponds to an unphysically fast dephasing rate. Note that the horizontal axis, *Broadening [meV]*, is not linear. Reproduced from Ref. [385] with permission from the American Chemical Society.

tion accuracy for small and large broadenings, but maintain comparative predictive powers across broadenings of 25 meV to 100 meV. Bayesian CNNs (blue trace) are the least robust with respect to changes in the broadening, with accurate predictions only within the 45 meV to 80 meV interval. Nevertheless, they demonstrate their predictive power for varying broadenings despite having been trained on peaks of one particular broadening, indicating that ML models can indeed be trained to identify relevant correlations without being overly sensitive to the broadening of the peaks.

4.3.4 ASSOCIATING MACHINE LEARNING MODELS WITH EXPERIMENTAL STUDIES

A long-standing goal of experimental materials characterization of organic optoelectronic materials is a map of how electronic properties are distributed in space as a result of different instantaneous molecular configurations. Scanning probe^{440–442} and super-resolution optical measurements⁴⁴³ can provide readouts of electronic properties on length scales below the diffraction limit. Simultaneously, single-particle measurements^{444,445} can provide a bottom-up understanding of how optoelectronic properties evolve from molecular precursors. For conductive polymers like PEDOT:PSS, single-particle measurements have been particularly difficult to employ due to the lack of emission in these materials caused by rapid quench-

ing. Simultaneously, single-particle measurements have tremendous spectroscopic utility due to reduced inhomogeneous broadening. Use of high quality-factor optical microresonators as the readout for ultrasensitive photothermal spectroscopy⁴⁴⁶ has allowed the first single-particle optical measurements to be performed on PEDOT:PSS,⁴³⁹ even down to a single or a small number of polymer strands. This study provided an experimental bound for the line broadening used in the above simulations. More recently, optical microresonator spectroscopy has been used to show how annealing processing act on single PEDOT:PSS polymer particles.⁴⁴⁷ A means of directly connecting spectral measurements on single PEDOT:PSS polymer strands and particles to electronic couplings would significantly amplify the information content of these experiments.

4.4 CONCLUSION

Our findings suggest that data-driven models can identify physical property-property correlations between the measurable electronic absorption spectra and the strength of intermolecular electronic couplings in organic electronics, which in turn determine the charge transport. While the presented ML models provide coupling strength estimates with limited accuracy, relative estimates with respect to reference coupling strengths show promising error rates. Using the trained Bayesian CNN model to classify given electronic absorption spectra above or below an *a priori* selected reference coupling strength displays an error rate of only 12.6%, and as low as 10% at high coupling regime. With such a promising error rate, we suggest using the trained models as classifiers to evaluate the performance of fabrication procedures and post-processing steps. Further investigations towards the construction of reliable and transferable ML models, notably the usage of ensemble methods such as AdaBoost,^{448,449} or mixture density networks,^{450,451} might allow for more detailed insights into the relation between couplings and electronic absorption spectra. Another important venue for improvement of our approach is the incorporation of features, such as structural information, which would introduce the notion of similarity between complexes. We believe that the combination of the developed strategies with spectroscopy techniques and its integration with

autonomous experimentation^{7,81,417,418} has the potential to enhance characterization and accelerate the optimization of organic materials. This study demonstrates that data-driven approaches enable the successful identification and quantification of proxy measurements, which has the potential to transform conventional experimentation strategies to promote scientific discovery. As experimental techniques for providing optical readouts improve in sensitivity, spatial resolution, and access to different spectral features, growth in theoretical treatments will allow one to draw deeper connections between these measurements and the underlying molecular structure. We also envision the use of spectroscopic methods to measure spectra of nanoaggregates with high-finesse toroidal optical cavities.

5

How machine learning can assist the interpretation of *ab initio* molecular dynamics simulations and conceptual understanding of chemistry

Apart from minor modifications, this chapter was originally published by the Royal Society of Chemistry as:

How machine learning can assist the interpretation of *ab initio* molecular dynamics simulations and conceptual understanding of chemistry. Florian Häse, Ignacio Fdez. Galván, Alán Aspuru-Guzik, Roland Lindh and Morgane Vacher. *Chem. Sci.* **10** (8), 2298–2307 (2019).

Reproduced from Ref. [72] with permission from the Royal Society of Chemistry.

ABSTRACT

Molecular dynamics simulations are often critical to the understanding of mechanisms, rates, and yields of chemical reactions, where experimental approaches cannot resolve the required length- or time-scales. However, one challenge to the deployment of molecular dynamics simulations is the in-depth analysis of large amounts of generated data, and the extraction of valuable insights to foster scientific understanding. In the present study, we suggest to employ emerging machine learning analysis tools to extract relevant information from simulation data without *a priori* knowledge. We demonstrate this approach on calculations of time-scales of the thermally activated decomposition of 1,2-dioxetane, which we estimate from *ab initio* molecular dynamics simulations and predict with machine learning models. Dissociation times predicted by the trained machine learning models are in good agreement

with the *ab initio* estimates. With this agreement, we analyze the trained models in detail and use them to probe the influence of specific thermal activations on the dissociation. We find that the trained models evidence empirical rules that are, today, part of the common knowledge in chemistry. This study, therefore, constitutes a step towards opening the way for conceptual breakthroughs in chemistry where machine analysis would provide a source of inspiration to humans.

5.1 THERMALLY ACTIVATED CHEMILUMINESCENCE

Computer simulations are a key complement to laboratory experiments for scientific discovery, especially when experiments are time-consuming, resource-demanding, or technically challenging (see Sec. 1.2). Simulations can provide insights into molecular processes that are inaccessible experimentally. For example, molecular dynamics (MD) simulations are essential to understanding the mechanisms, rates and yields of chemical reactions by studying the time evolution of matter at the atomic level. *Ab initio* MD allows us to compute the time evolution of a many-particle system at discrete time steps and propagate the system *via* sophisticated integration methods. At each time step, the energies and forces exerted on the nuclei are calculated on-the-fly from the output of an electronic structure method. With the growing complexity of commonly studied chemical systems and the increasing demand for improved accuracy, the resource requirements for MD simulations steadily increase. Typical time and length scales that are accessible with *ab initio* MD range from hundreds of femtoseconds to tens of picoseconds for systems consisting of tens to a few hundreds of atoms. As simulations grow in complexity, their practicality for guiding and understanding molecular systems on the atomic level may become obscured. Simple lessons are easily lost among gigabytes or terabytes of data. In the present work, we propose to use data-driven methods to streamline the analysis of MD simulations and enhance their interpretability. Ultimately, we hope to see machines and artificial agents providing a source of inspiration to human researchers to conceptualize scientific findings and elaborate novel insights in chemistry, thus aiding in the hypothesizing step of the scientific method (see Sec. 1.1). This fundamental

challenge has been recognized as one of the six open questions for the simulation of matter in the 21st century⁴⁵², and the work presented in this chapter demonstrates a step towards achieving this goal.

We choose to study the timescale of the thermally activated chemiluminescent decomposition of 1,2-dioxetane as a test application (see Fig. 5.1a). Chemiluminescence describes the emission of light as a result of a chemical reaction. This process is called bioluminescence when occurring in living organisms, as it is the case in the example of the firefly. Chemiluminescence, bioluminescence, and approaches to study these processes computationally have recently been reviewed in detail.^{453,454} Both phenomena are increasingly used in biological and chemical analysis methods⁴⁵⁵ across several fields such as DNA sequencing,⁴⁵⁶ immunoassays as an alternative to radioactive isotopes⁴⁵⁷ and as a sensitive probe for mechanical stimulations.^{458,459} Yet, the applicability of chemiluminescence as an analytic tool is limited by the light emission efficiency, *i.e.*, the chemiluminescence yield. The smallest compound with known chemiluminescent properties is 1,2-dioxetane. Upon thermal activation, it decomposes into two formaldehyde molecules in a two-step process: (i) first, the oxygen-oxygen bond breaks to form a biradical, and (ii) then the carbon-carbon bond dissociates (see Fig. 5.1a).^{460,461} Non-adiabatic transitions from the electronic ground state to electronic excited states during the decomposition reaction lead to chemiexcitation. The resulting electronic excited states can relax back to the electronic ground state by emitting light in the form of photons. Previous *ab initio* MD simulations on methyl-substituted dioxetane molecules suggest that the dissociation timescale determines the chemiexcitation yield:^{462,463} the longer the molecule stays in an *entropic trap* before dissociating, the higher the chemiexcitation yield. Estimating and understanding the dissociation times of dioxetanes are thus essential for rationalizing chemiluminescence yields measured experimentally, which is a critical challenge to the design of more efficient chemiluminescent compounds. However, it is *a priori* unclear which molecular modifications would eventually result in substantial increases in the dissociation time. Testing possible hypotheses with *ab initio* methods is an arduous process due to the computational demand and can easily exceed

available computing resources.

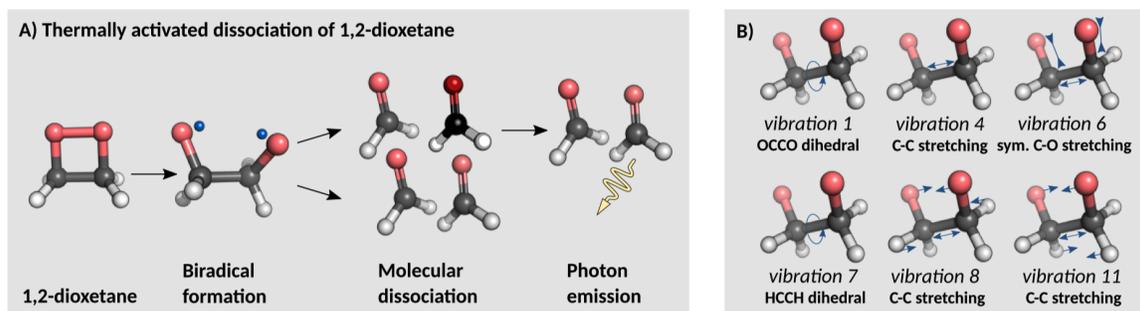


Figure 5.1: Thermally activated dissociation of the 1,2-dioxetane molecule. (A) Dark and chemiluminescent dissociation reactions. (B) Schematic representation of the relevant normal modes. Normal mode 1 corresponds mainly to the OCCO dihedral angle while normal mode 7 corresponds mainly to the HCCH dihedral angles.

In the present work, we illustrate how machine learning (ML) models can be used to approach this task. Specifically, we employ Bayesian neural networks (BNNs), which are trained to predict the dissociation times of 1,2-dioxetane from either only initial nuclear positions or initial nuclear positions and momenta. We demonstrate that the trained BNNs can spark scientific insights in two different ways: (i) by rapidly testing already formulated hypotheses, and (ii) by autonomously identifying and highlighting relevant physical correlations. We apply both approaches to the unmethylated and tetramethylated dioxetanes, the latter presenting longer dissociation timescales and thus higher chemiexcitation yields. Despite the limited size of the dataset, consisting of 250 trajectories simulating only 250 fs at time steps of 0.24 fs, the trained models achieve high prediction accuracies. This observation by itself presents an advancement over prior studies where datasets comprised orders of magnitude more data points.^{372,373} More importantly, we illustrate how the ML models themselves and the changes they have undergone during training can be interpreted to inspire physical insights. We further use the trained BNNs to predict dissociation times of vibrationally excited states of the unsubstituted 1,2-dioxetane. A detailed understanding of the effect of specific nuclear distortions, which could be modulated *via* chemical modifications, enables opportunities for targeted molecular design. Aiming at gaining physical insights, we demonstrate that training BNNs to reproduce the dissociation times of 1,2-dioxetane with high accuracy structures the BNNs such that they can evidence chemical rules which

connect nuclear positions (and velocities) with the associated physical dissociation times.

Many efforts have recently been devoted to the construction of efficient ML models which *learn* and predict potential energy surfaces orders of magnitudes faster than electronic structure calculations, in an attempt to reduce the computational cost of MD simulations.^{330,375,464–469} For example, ML models have been used to reproduce energies at the level of hybrid density-functional theory from lower-level calculations and molecular descriptors^{470,471} or directly from the Cartesian coordinates and the nuclear charges.^{371,472,473} The present work targets a higher level of abstraction and proposes to use ML models to directly predict a specific outcome of the MD simulation, thus bypassing the construction of potential energy surfaces and avoiding the computation of the time evolution of the studied system. However, we focus neither on reducing the computational cost of electronic structure calculations in *ab initio* MD simulations, nor on describing chemiluminescent or bioluminescent reactions with the best possible quantitative accuracy. Instead, we train ML models on already simulated trajectories of a given chemical reaction and, at a pre-defined level of theory, interpret and conceptualize physical insights into the studied reaction *via* the trained ML models. The strategies that we outline in this work to interpret MD simulations are agnostic to model-specific information and are thus expected to generalize to other model systems.

5.2 DATA-DRIVEN CHEMILUMINESCENCE

We proceed by detailing the individual steps of our approach to provoking physical insights from MD simulations using data-driven strategies. We will first detail the *ab initio* MD simulations from which we generated the dissociation trajectories and then outline how BNN were trained to reproduce the dissociation times of dioxetane.

5.2.1 AB INITIO MOLECULAR DYNAMICS SIMULATIONS

We set up the *ab initio* MD simulations similarly to previously published works.^{454,463} The electronic structure of dioxetane was modeled with the complete active space self-consistent field (CASSCF) method^{474,475} where we average over the four lowest energy singlet states equally. The chosen active space consisted of 12 electrons and ten orbitals: the four σ and four σ^* orbitals of the four-membered ring, plus the two oxygen lone-pair orbitals perpendicular to the ring. We used the ANO-RCC basis set with polarized triple-zeta contraction.^{476,477} The system is propagated in time according to Born-Oppenheimer dynamics with a time step of 10 au (~ 0.24 fs) and taking into account all nuclear coordinates. Only the electronic ground state is included in the simulations, and non-adiabatic transitions to electronic excited states are not allowed. The implementation of this approach is provided in the OpenMolcas package.⁴⁷⁸ Individual trajectories are initialized and propagated from the transition state for the oxygen-oxygen bond breaking (see Fig. 5.1a), which controls the overall reaction rate. As suggested in previous theoretical studies of post-transition state dynamics,⁴⁷⁹ we add a small amount of kinetic energy, 1 kcal/mol, towards the biradical region where the oxygen-oxygen bond is broken. Initial positions and momenta for the 250 trajectories along all normal modes were sampled from a Wigner distribution using the Newton-X package to reproduce the vibrational ground state. The normal modes were calculated at the transition state using the electronic structure method mentioned above. Dissociation is considered to occur when the carbon-carbon bond length exceeds 2.4 \AA , which amounts to two times the van der Waals radius of a carbon atom. Choosing a slightly smaller or larger value for the bond length dissociation threshold has been shown to not change any relative comparisons, nor the findings regarding the entropic trap and the effect of the singlet excited states.

5.2.2 MACHINE LEARNING PREDICTIONS

We estimate dissociation times of dioxetane with two probabilistic ML models implemented as feedforward fully connected BNNs. Conventional neural networks are constructed as a set of nodes with connections between them (see Fig. 5.2). Each neuron in a conventional neural network is characterized by a weight, \mathbf{w} , a bias, \mathbf{b} , and an activation function, f_{act} . For a given input, \mathbf{x} , a single neuron performs the operation

$$\mathbf{y} = f_{\text{act}}(\mathbf{w}\mathbf{x} + \mathbf{b}), \quad (5.1)$$

where \mathbf{y} denotes the output of the neuron. The architecture of a neural network is defined by a set of hyperparameters which determine, for instance, the number of layers, the number of neurons per layer and the activation function of the neurons. In the case of BNNs, both weights, \mathbf{w} , and biases, \mathbf{b} , are modeled as random variables and sampled from probability distributions which are conditioned either on model parameters in preceding layers or the input parameters (see Fig. 5.2a) The BNN therefore models an output distribution which can be adapted to resemble a desired distribution of targets by adjusting the distributions of weights and biases for each individual neuron *via* Bayesian inference (see Sec. 2.2.2). As such, BNNs propagate information based on entire parameter distributions as opposed to conventional neural networks which compute a target based on single valued parameters (see Fig. 5.2b). BNN thus can retain the flexibility of conventional neural networks but provide a more robust framework for the identification of relevant correlations between inputs and outputs, especially for small and medium sized datasets.⁴⁸⁰

We construct two different BNN models for different sets of input features. The first model, hereafter noted bnn1, uses only the initial nuclear geometry of the molecule. The second model, hereafter noted bnn2, also uses the initial nuclear velocities in addition to the geometry. The nuclear geometries and velocities are encoded in normal mode coordinates to account for translational and rotational invariances. The normal modes on which coordinates were projected are calculated for the transition state structure, as indicated above. For the

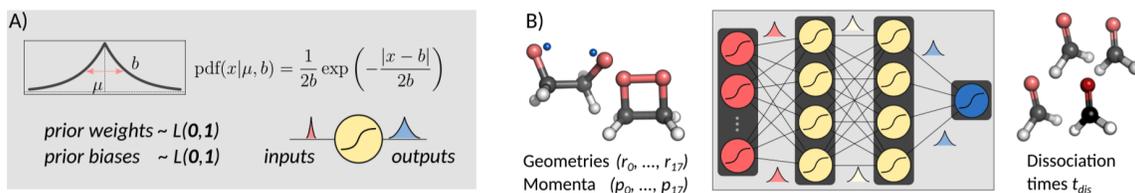


Figure 5.2: Illustration of a Bayesian Neural Network (BNN). (A): A Bayesian neuron defines a mathematical operation based on an activation function, a distribution of weights w and a distribution of biases b intrinsic to the neuron. Every input x is processed by sampling one instance of weights and biases from the distributions and applying the activation function. (B) A BNN consists of a set of interconnected Bayesian neurons. The neurons in the network are organized in layers and can differ in their activation functions, their weight distributions, and their bias distributions. The two models constructed in the present work use different sets of input features. The first model, hereafter noted *bnn1*, uses only the initial nuclear geometry of the molecule. The second model, hereafter noted *bnn2*, also uses the initial nuclear velocities in addition to the geometry. The nuclear geometry and velocities are given in normal mode coordinates to account for translational and rotational invariances (see main text). For the unmethylated dioxetane molecule consisting of 8 atoms, *bnn1* has, therefore, 18 input features, while *bnn2* has 36 input features. For the tetramethylated dioxetane molecule consisting of 20 atoms, *bnn1* and *bnn2* have thus 54 and 108 input features, respectively.

unmethylated dioxetane molecule consisting of 8 atoms, *bnn1* is modeled with 18 input features, while *bnn2* is modeled with 36 input features. For the tetramethylated dioxetane molecule consisting of 20 atoms, *bnn1* (*bnn2*) has 54 (108) input features.

Datasets to train the BNN models were extracted from the *ab initio* MD simulations by sampling every fifth frame up to successful dissociation along each of the 250 trajectories, amounting to a total of 11,650 selected frames and their corresponding dissociation times. We split the total dataset into a training set (80%, 9,320 frames), a validation set (10%, 1,165 frames) and a test set (10%, 1,165 frames). Test data were selected as random samples from the entire set and were not used for any part of the training procedure other than for reporting the out-of-sample prediction accuracy after all BNN models had been trained. An informative and diverse training set was assembled from the remaining 90% of the dataset *via* principal component analysis (PCA), similar to the training set selection in Chapter 3.⁴⁸¹ We selected the frames for the training set, which are maximally separated in the reduced PCA space spanned by the most contributing principal components. This protocol has shown to improve prediction accuracies in the context of excitation transfer property predictions.³³⁰ The frames, which were not selected with this protocol, were used as the validation set.

The BNN models for the present study were implemented in the probabilistic programming

library EDWARD,⁴⁸² and model parameters were updated via variational inference using the ADAM optimization algorithm.³⁷⁹ We parameterized the distributions of weights and biases as Laplace distributions. This choice is made in order to construct interpretable models. While the training set is used to optimize the model parameters, \mathbf{w} and \mathbf{b} , of a BNN model of a given architecture, the validation set serves as a benchmark to determine the best performing architecture. For both bnn1 and bnn2, we conducted extensive hyperparameter searches to determine the best performing BNN models. The most accurate BNN model was selected as the model with the lowest prediction error on the validation set. For both bnn1 and bnn2, we conducted extensive hyperparameter searches to determine the best performing BNN models. The most accurate BNN model was selected as the model with the lowest prediction error on the validation set. Details on the implementation of the BNN models and the hyperparameter optimization including the scanned hyperparameters and obtained performances of different BNN architectures, are reported in the supporting information of Ref.⁷² Despite different numbers of input features, we found that the best performing hyperparameters for bnn1 and bnn2 were mostly similar. Both BNN models were trained to learn the dissociation time of 1,2-dioxetane and tetramethyl-1,2-dioxetane from only geometries (and velocities) without additional *a priori* knowledge about the dynamics of the chemical reaction of interest.

5.3 DERIVING MECHANISTIC INSIGHTS FROM DATA-DRIVEN FINDINGS

In the following, we use the trained ML models to study the mechanism of the dissociative reaction of 1,2-dioxetane. The BNNs trained on the unmethylated dioxetane molecule is analyzed in detail. Unless stated otherwise, results are reported for this compound, and not the tetramethylated variant. We begin our analysis by validating that the trained BNNs are indeed capable of predicting dissociation times within reasonable accuracy. Then, we proceed with analyzing the architecture of the trained BNNs models to identify nuclear coordinates relevant to the dissociation timescale. Finally, we use the BNNs to predict dissociation times for vibrational excitations to test hypotheses about physical correlations.

5.3.1 VALIDATION OF THE DISSOCIATION TIME PREDICTIONS

First, we present the comparison of the carbon-carbon bond dissociation times obtained from the *ab initio* MD simulations with the dissociation times predicted by the bnn1 and bnn2 models. Results for these comparisons are illustrated in Fig. 5.3. Uncertainties in the predicted dissociation times were quantified *via* the standard deviations of the predicted dissociation time distributions. We find that BNNs can predict dissociation times with high accuracies, despite the medium-sized dataset used for training (see Sec. 5.2.2) A detailed analysis of the sampling efficiency, *i.e.*, the achieved performance for different training set sizes is reported in the appendix of Ref. [72]. We find that the prediction accuracy is improved when supplementing the molecular geometry with the velocities of the atoms (see Fig. 5.3b): we find a mean absolute deviation (MAD) of 2.40 fs for bnn2 and 6.55 fs for bnn1. This observation, combined with the observation that prediction accuracies on the validation set are similar for bnn1 or bnn2, indicates that the velocities are indeed relevant to the dissociation times of the molecule. The BNN models also generally predict dissociation times with higher uncertainties when the deviations between the predicted and the true dissociation times are larger. We note that the prediction accuracy can likely be improved with strategies such as cross-fold validation for an effective increase of the training set size as proposed in other studies,^{370,483} or testing different models such as kernel-based methods.^{471,484,485} However, in this study, we aim to focus on the interpretability of the trained models and $r^2 = 0.97$ for bnn2 is considered to be sufficient for this purpose.

Once trained, ML models can be extensively used to inexpensively query dissociation times with high accuracy at low computational costs. As a reference, we found that one processor can generate a set of 250 initial conditions within less than a minute, and predict the corresponding dissociation times with the trained BNN models within a few hours, while simulating a trajectory from *ab initio* MD requires approximately 31.5 hours for 160 fs of simulation time. We therefore consider the trained BNN models as a result by themselves and analyze them to gain physical insights. In particular, we are interested in understanding the

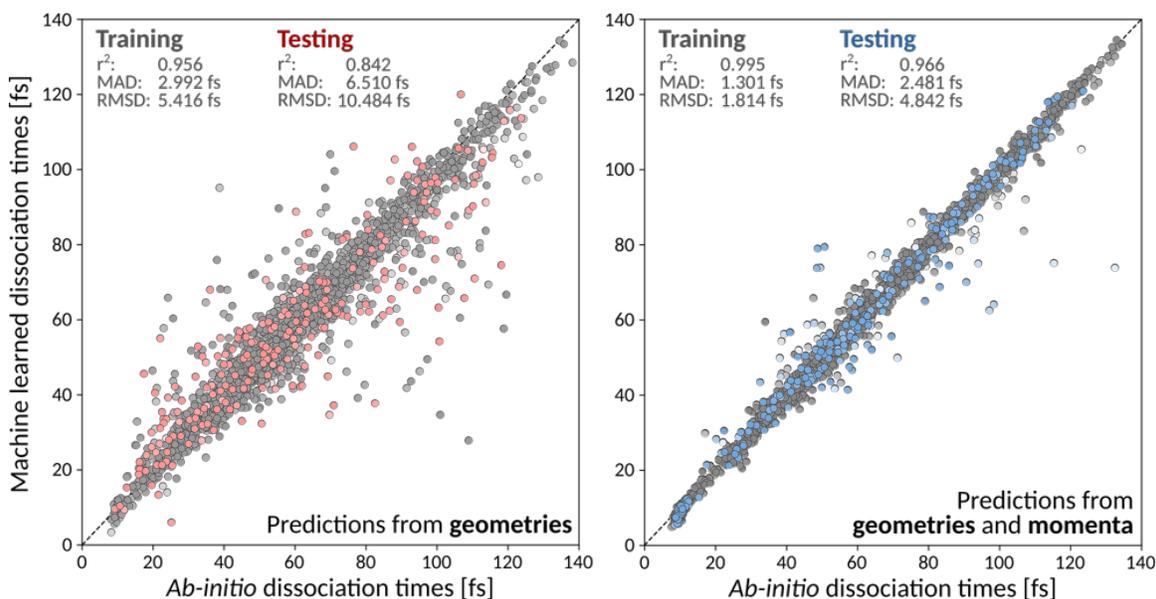


Figure 5.3: Out-of-sample predictions for best performing bnn1 model (left panel) and bnn2 model (right panel). For all frames in the test set (1165 frames), we depict the dissociation times predicted by the BNN models in comparison to the true dissociation times. Root mean square deviations (RMSDs), mean absolute deviations (MADs), and coefficients of determination (r^2) are reported. Depicted points are colored based on the predicted uncertainty of the BNN model. The grey dashed line indicates perfect agreement between predicted and true dissociation times.

following aspects: How does the BNN reproduce the carbon-carbon dissociation timescales with such a good agreement? What correlations has the BNN identified to achieve high accuracies?

5.3.2 ANALYSIS OF THE TRAINED MODELS TO FIND CORRELATIONS

We use the Laplace distributions as a prior for the distributions of weights and biases in the BNN models, as it facilitates efficient pruning in the model parameters,⁴⁸⁶ and is the equivalent to L1 regularization in the Bayesian context.⁴⁸⁷ This choice is expected to enable the BNN models to highlight only the normal modes among the inputs, which are most relevant for accurate predictions of dissociation times. Generally, we consider normal modes to be more influential on the predicted dissociation times if the modes of the coefficient distributions that process the associated features in the input layer are larger. Normal modes that are relevant to the accurate prediction of dissociation times are also expected

to be relevant to the underlying physical process. Fig. 5.4 illustrates the distributions of coefficient magnitudes for displacements along individual normal modes for the fully trained network architectures, bnn1 and bnn2. Normal modes are numbered from 0 to 17, where 0 corresponds to the reaction coordinate. Cartesian coordinates projected onto the normal modes are denoted with \mathbf{r}_i , while projected velocity components are denoted with \mathbf{v}_i . Input features that are processed by coefficients with generally large magnitude are expected to modulate the dissociation time scale. However, this analysis does not allow us to conclude the nature of the modulation, *i.e.*, whether excitations along this normal mode generally promote or delay dissociation. The choice of normal modes as input coordinates reflects the intention to interpret the trained model architectures to gain physical insights, and probing other input features is beyond the scope of the present work.

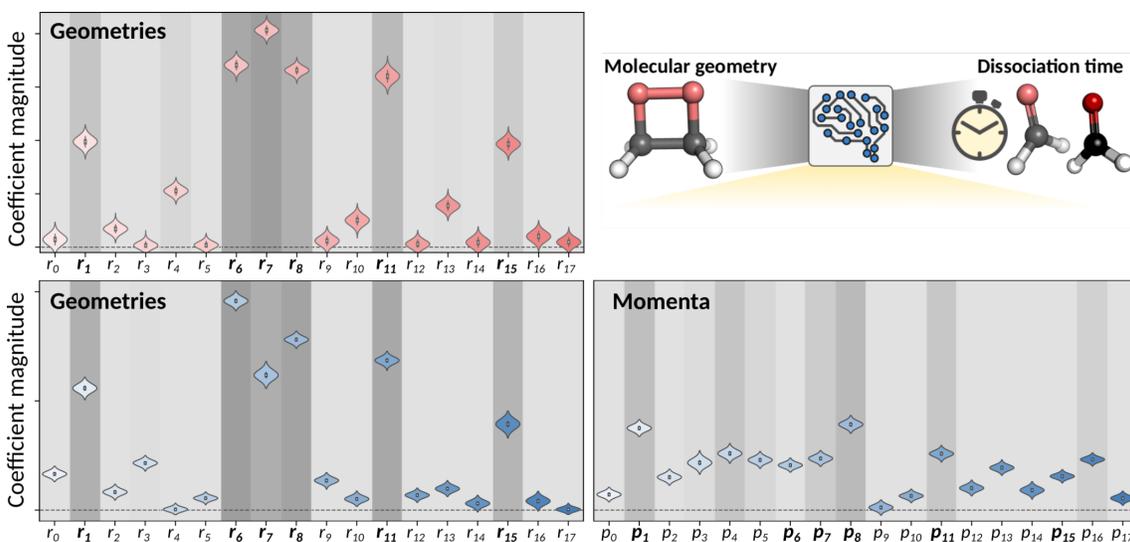


Figure 5.4: Coefficient magnitude distributions of the input features of the trained bnn1 (upper panel) and bnn2 (lower panels). Input features $\{r_i\}$ and $\{v_i\}$ correspond to geometry coordinates and velocities along normal modes. The numbering of the normal modes goes from 0 to 17, 0 representing the reaction coordinate.

Among the 18 normal modes, bnn1 identifies the modes 6, 7, 8, and 11 to be most influential to the decomposition reaction. All four of these modes are illustrated schematically in Fig. 5.1b. Normal mode 6 corresponds to the symmetrical stretching of the two carbon-oxygen bonds in combination with the out-of-phase stretching of the central carbon-carbon bond. Normal mode 7 represents to motions around the hydrogen-carbon-carbon-hydrogen

dihedral angles. Normal modes 8 and 11 constitute the stretching of the central carbon-carbon bond in phase with symmetric out-of-plane motions of the two formaldehyde moieties. In bnn2, the coefficients of the initial nuclear geometry (\mathbf{r}_0 to \mathbf{r}_{17}) are generally larger in magnitude than those of the initial velocities (\mathbf{v}_0 to \mathbf{v}_{17}). Similar to bnn1, distortions along \mathbf{r}_6 , \mathbf{r}_7 , \mathbf{r}_8 and \mathbf{r}_{11} are recognized as relevant. In addition, the perceived influence of \mathbf{r}_1 seems to be more expressed in bnn2 than in bnn1. Normal mode 1 corresponds to motions around the oxygen-carbon-carbon-oxygen dihedral angle (see Fig. 5.1b). Among the remaining 18 inputs corresponding to the initial velocities, the features \mathbf{v}_1 and \mathbf{v}_8 present the largest coefficient magnitudes. Normal modes 1 and 8 are already identified as important according to the coefficients of the initial position features.

5.3.3 USE OF THE TRAINED MODELS TO TEST HYPOTHESES

The analysis presented in Sec. 5.3.2 highlighted the prevalent motions which determine the dissociation of dioxetane from a region in phase space close to the biradical transition state. However, the nature of the modulation induced by the relevant normal modes is not revealed yet. We suggest to go one step further and use the trained BNN models to predict dissociation timescales for 17 ensembles of 250 initial conditions, each ensemble representing a vibrational state that is excited to the first level along one particular normal mode. For example, the third ensemble exclusively contains vibrational states which are excited along normal mode 3, while it remains in the ground state with respect to all other normal modes. Ensemble 0 represents the reference vibrational ground state for all normal modes but contains initial conditions different from those on which the BNN models have been trained. For each ensemble, we predict the dissociation times for all 250 initial conditions and determine the dissociation half-time, *i.e.*, the time after which half of the trajectories have successfully dissociated. The predicted dissociation half-times are illustrated in Fig. 5.5a. We find that the vibrational excitations can alter the dissociation times from 49 fs to 63 fs while the reference dissociation time of the vibrational ground state is at 60 fs. We observe that vibrational excitations along the normal modes mostly tend to accelerate the decom-

position reaction, which agrees with the expectation that the addition of kinetic energy to the system generally destabilizes the molecule. Excitations of normal modes 4, 6, 8, and 11 appear to accelerate the dissociation the most, while excitations along normal mode 3 seem to stabilize the molecule on average and delay the dissociation. Normal modes 6, 8, and 11 were already classified as relevant in the previous analysis. Normal mode 3 corresponds to the antisymmetric stretching of the two carbon-oxygen bonds and is thus directly antagonistic to normal mode 6. Our observations suggest that for the carbon-carbon bond to break earlier rather than later, the two formaldehyde moieties need to be planarized simultaneously to the symmetric shortening of the carbon-oxygen bonds, while an antisymmetric stretching of the carbon-oxygen bonds delays the dissociation. High-frequency modes involving the hydrogen atoms, in contrast, seem to only moderately affect the dissociation timescales.

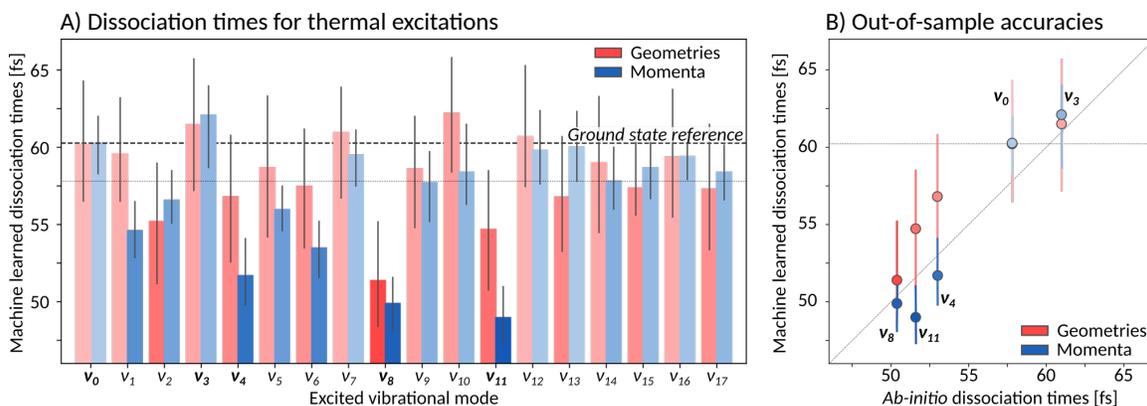


Figure 5.5: Dissociation half-times for 17 ensembles of 250 initial conditions representing different vibrationally excited states. Ensemble 0 is the reference vibrational ground state. Otherwise, ensemble n corresponds to a vibrational excitation along normal mode n , and ground state along other normal modes. (A) Predicted dissociation half-times by the two trained BNN models. The solid and dashed horizontal lines indicate the dissociation time for the ensemble 0, predicted and extracted from the simulations, respectively. (B) Comparison of the dissociation half-times predicted by the BNN models and obtained from *ab initio* MD simulations, for five ensembles of 250 trajectories. The error bars represent the 95 % confidence intervals of the predictions.

Although normal mode 7 was indicated as being relevant for the accurate prediction of dissociation times in Fig. 5.4, vibrational excitations along only this mode seem to overall have no significant effect on the dissociation half-time compared to the reference dissociation half-time of the ground state (see Fig. 5.5a). To further investigate this observation, we used the trained `bnn2` to predict the dissociation times for ensembles of initial conditions, where

each ensemble represented a vibrational state that is excited to the first level along two particular normal modes to probe for any cooperative second-order effects. The predicted dissociation half-times obtained for each pair of normal mode excitations are reported in the appendix of Ref. [72]. We find that the simultaneous vibrational excitation of normal mode 7 with any other normal mode delays the dissociation. This observation indicates that the relevance of an excitation along the seventh mode is not based on its direct impact on the dissociation timescale, but instead because it prevents accelerations of the dissociation which would be induced by simultaneous excitations along other normal modes. This result is, therefore, in agreement with the large coefficient magnitudes for the seventh mode presented in Fig. 5.4a and also highlights the complementary nature of the two proposed analysis strategies.

5.3.4 AB INITIO MOLECULAR DYNAMICS SIMULATIONS OF VIBRATIONALLY EXCITED STATES

We probe the accuracy of the dissociation times for vibrationally excited ensembles predicted by the BNN models (see Fig. 5.5a) by simulating the *ab initio* MD of four selected ensembles as well as the reference ensemble for the vibrational ground state. Specifically, we focus on the ensembles with vibrational excitations along normal modes 3, which delays the dissociation half-time, and 4, 8, and 11, which generally accelerate the dissociation. Indeed, excitations of the fourth normal mode, which corresponds to the stretching of the carbon-carbon bond, could be seen as the naïve choice for promoting dissociation without a chemical foundation. Comparisons of predicted and calculated half-times are illustrated in Fig. 5.5b. We find that the qualitative trend is indeed correctly predicted by the BNN models. Overall, the prediction uncertainties are within the same order of magnitude as for the reference ensemble, although the excited normal modes present with more substantial initial nuclear distortions that may not be present in the training set. Generally, the predictions of bnn2 are in better agreement with the ground truth with a root mean square deviation (RMSD) between predicted and targeted values being about 1.6 times smaller than for the predictions of bnn1. In fact, bnn1 tends to overestimate dissociation times. The BNN predictions also

Table 5.1: Average number of frustrated dissociations per trajectory for each ensemble of 250 trajectories

Ensemble	0	3	4	8	11
n_{frust}	1.68	1.52	1.18	1.08	1.17

partially validate the naïve thought that excitations of the carbon-carbon bond stretching promote dissociation, as we find dissociation times for excitations of the fourth mode to be about 5 fs faster than the dissociation times of the ground state ensemble. However, the BNN models also successfully identify other nuclear modes, which decrease the dissociation times even further.

Previous theoretical studies explained modulations on the dissociation times among the 1,2-dioxetane trajectories by the presence of *frustrated* dissociations,⁴⁶² *i.e.*, significant stretching of the central carbon-carbon bond followed by a shortening rather than a *successful* breaking of the bond. Frustrated dissociations thus postpone the final successful dissociation. We further compute the average number of frustrated dissociations per trajectory for each ensemble of 250 trajectories (see Tab. 5.1). Generally, the vibrationally excited ensembles present fewer frustrated dissociations than the reference ground state ensemble since more kinetic energy is available to cross the transition barrier. We also observe that fewer frustrated dissociations are associated with shortened average dissociation half-times. This observation is particularly apparent for ensemble 8, which exhibits the most substantial decrease in the average number of frustrated dissociations (35 %) and the largest decrease in the dissociation time (13 %). The BNN have thus identified the nuclear coordinates that overall accelerate dissociations by reducing the number of frustrated dissociations. However, ensemble 3 presents a lower average number of frustrated dissociations compared to the reference ensemble despite an extended dissociation half-time. This observation indicates that the number of frustrated dissociations is not the only determining factor for modulated dissociations.

5.4 DISCUSSION

We start our discussion on the capabilities of the trained ML models and their practicality to streamline the analysis of MD trajectories with the goal to spark physical insights and promote scientific intuition by interpreting the identified correlations (see Sec. 5.3) in the context of chemistry. We further proceed by outlining the implications of our observations for chemiexcitation and chemical design on the example of the tetramethylated dioxetane molecule.

5.4.1 INTERPRETATION OF THE FINDINGS OF THE TRAINED MODELS

The BNN models suggest that the carbon-carbon bond breaking must be associated with the formation of a shorter carbon-oxygen bond and the planarity of the formaldehyde moieties, and confirm that a naïve excitation of only the carbon-carbon bond stretching is not sufficient for rapid dissociations. These insights provide evidence for at least three concepts which are fundamental to chemistry: the octet rule,^{488–490} the relation between bond order and bond length^{491,492} and orbital hybridisation.⁴⁹³ The octet rule states that atoms with an atomic number $Z \geq 5$ favor electronic configurations where they combine in such a way that each atom has a full valence shell occupied by eight electrons. This electronic configuration, naturally assumed by noble gases, is associated to maximal stability in molecular systems. The carbon atoms in dioxetane which have only four electrons in their valence shells will therefore favor the formation of four covalent bonds to acquire the four missing electrons. As the carbon-carbon bond breaks, each carbon atom needs to form a new bond with one of the remaining neighboring atoms to keep the stable electronic configuration intact, which is achieved by forming a double bond with the oxygen atoms. The number of bonding electrons between two atoms, quantified as the bond order, determines the length of a bond. Transforming the single bond between the carbon and oxygen atoms into a double bond requires a shortening of the carbon-oxygen bond, which is represented by normal mode 6. The change in the molecular shape, observed in the form of a planarization of the

formaldehyde moieties, can be rationalized by a change in the orbital hybridization, *i.e.*, the mixing of the $2s$ and $2p$ atomic orbitals, to achieve a minimal repulsion energy between pairs of electrons. The molecular shape can be rationalized with the valence shell electron pair repulsion (VSEPR) model.^{494,495} As long as each carbon atom is surrounded by four atoms, as it is the case in the biradical transition state before the dissociation, the $2s$ and the three $2p$ orbitals mix in a tetrahedral arrangement to form the four covalent bonds, which is generally referred to as sp^3 hybridization. When the carbon-carbon bond breaks and each carbon atom is only surrounded by three other atoms, the $2s$ orbitals mix only with two of the $2p$ orbitals to form sp^2 hybridized orbitals while the remaining $2p$ orbital remains unchanged to form the π bond by parallel overlap. The repulsion energy in this new configuration is minimized by a trigonal planar geometry. Upon the breaking of the carbon-carbon bond, the two formaldehyde moieties thus planarize, which corresponds to the motions observed in normal modes 8 and 11.

5.4.2 IMPLICATIONS FOR CHEMIEXCITATION AND CHEMICAL DESIGN

Since the dissociation timescale determines the chemiexcitation yield,⁴⁶³ more efficient chemi-luminescent molecular systems could be designed by identifying the nuclear coordinates which reliably delay the dissociation dynamics. The BNN models identified normal modes 3 and 8 as the nuclear coordinates which induce the largest deviations from the reference ensemble (see Fig. 5.5b). The chemiexcitation yield in 1,2-dioxetane was measured to be about 0.3%.⁴⁹⁶ A chemical substitution that would promote the simultaneous planarization of the two formaldehyde moieties together with the stretching of the central carbon-carbin bond would accelerate the dissociation and therefore decrease the chemiexcitation yield. However, a chemical substitution that induces an asymmetric stretching of the two carbon-oxygen bonds would delay the dissociation and therefore increase the chemiexcitation yield. According to the kinetic model developed previously,⁴⁶³ delaying the dissociation time of the unmethylated 1,2-dioxetane by 3.2 fs, as observed for ensemble 3, is expected to increase the chemiexcitation yield by more than a factor of six.

Experimental evidence suggests that the chemiexcitation yield can be increased by approximately two orders of magnitude when substituting the four hydrogen atoms with four methyl groups.⁴⁹⁶ Initially, it was suggested that the chemiexcitation yield was due to an increase of the depth of the entropic trap induced by the greater number of nuclear coordinates in the methyl-substituted compound.⁴⁶⁰ Recently, the *ab initio* MD of the methyl-substituted dioxetane has been simulated.⁴⁶³ This later work suggested that the increase in the chemiluminescent yield can instead be attributed to the higher mass of the molecule. When training the BNN models on an ensemble of trajectories of the tetramethylated dioxetane, we found that the nuclear coordinates which presented with large coefficient magnitudes again correspond to a stretching of the carbon-oxygen bonds, a stretching of the central carbon-carbon bond, a planarization of the two ketone moieties and the oxygen-carbon-carbon-oxygen dihedral angle. These coordinates are similar to the normal modes 1, 4, 6, 8, and 11 in the unsubstituted 1,2-dioxetane molecule. This result suggests a similarity in the dissociation dynamics of the unsubstituted and methyl-substituted molecules and therefore illustrates that the significance of the methyl groups for increased chemiexcitation yield is more founded in the heavier mass than the additional number of degrees of freedom.

5.5 CONCLUSION

In this study, we constructed BNN models for the prediction of dissociation times of the unmethylated and tetramethylated 1,2-dioxetane molecules from *ab initio* MD from initial nuclear geometries and velocities. Despite the medium size of the dataset used for training, we obtain a sufficiently high prediction accuracy. We have further demonstrated that the BNN models trained in this study can highlight physically relevant trends in the generated data, based on which important fundamental concepts in chemistry are evidenced. The trained BNN models correctly predict that the naïve approach to promoting the dissociation by exciting the stretching of the central carbon-carbon bond is not the most efficient. Instead, the indications of the BNN models suggest that rapid dissociation requires a planarization of the two formaldehyde moieties and the simultaneous symmetric shortening of the carbon-

oxygen bonds. This observation can be connected to the octet rule, the relation between bond order and bond length, and to orbital hybridization, which are all rules that are part of today's common knowledge in chemistry. The present work therefore constitutes a step towards achieving one of the grand challenges in the 21st century⁴⁵² and opens thus the way for breakthroughs in chemistry where humans, inspired by the findings of machines, would develop new concepts.

Part II

Algorithms for closed-loop experimentation

This page is intentionally left blank.

6

Phoenix: A Bayesian optimizer for Chemistry

Apart from minor modifications, this chapter was originally published by the American Chemical Society:

Florian Häse, Loïc M. Roch, Christoph Kreisbeck and Alán Aspuru-Guzik. Phoenix: A Bayesian optimizer for chemistry. *ACS Cent. Sci.* **4** (9), 1134–1145 (2018).

Reproduced with permission from Ref. [233] Copyright 2018 American Chemical Society

ABSTRACT

In this chapter, we report PHOENIX, a probabilistic global optimization algorithm to identify the set of conditions of an experimental or computational procedure that satisfies desired targets. PHOENIX combines ideas from Bayesian optimization with concepts from Bayesian kernel density estimation. As such, PHOENIX allows tackling typical optimization tasks in chemistry for which objective evaluations are limited, due to either budgeted resources or time-consuming processes. PHOENIX proposes new conditions based on all previous observations, avoiding, thus, redundant evaluations to locate the optimal conditions. It enables an efficient parallel search based on intuitive sampling strategies implicitly biasing towards an exploration or exploitation of the search space. Our benchmarks indicate that PHOENIX is less sensitive to the response surface than established optimization algorithms. We showcase the applicability of PHOENIX on the *Oregonator*, a complex case-study describing a non-linear chemical reaction network. Despite the large search space, PHOENIX quickly identifies the conditions which yield the desired dynamic behavior. Overall, we recommend PHOENIX for the rapid optimization of unknown expensive-to-evaluate objective functions, such as experimentation or long-lasting computations.

6.1 DATA-DRIVEN STRATEGIES TO EXPERIMENT PLANNING

Optimization problems are ubiquitous in a variety of disciplines ranging from science to engineering. They can take various facets: finding the lowest energy state of a physical system, searching for the optimal set of conditions to improve experimental procedures, or identifying the best strategies to realize industrial processes. More generally, scientific discovery itself can be formulated as an optimization problem (see Sec. 2.3). For example, conditions for chemical reactions are optimized with systematic methods such as design of experiments (DoE).^{188–190} More recently, optimization procedures assisted chemists in finding chemical derivatives of given molecules to best treat a given disease,²³⁰ designing candidates for organic photovoltaics,⁴⁹⁷ organic synthesis,^{192,498} predicting reaction paths,^{499–501} and in automated experimentation.^{203,502–504} Often, these applications are subject to multiple local optima and involve costly evaluations of proposed conditions in terms of required experimentation or extensive computations. Optimization tasks are typically formulated with an objective function, which for a given set of parameters returns a measure for the associated merit. This task relates, for example, to measuring the yield of a chemical reaction conducted under specific experimental conditions. In the past, a variety of optimization algorithms have been developed. Gradient-based algorithms such as gradient descent,¹⁹⁴ conjugate gradient,¹⁹⁵ or the more sophisticated BFGS,^{196,197} are efficient at finding local optima. However, they require numerous evaluations, *i.e.*, conducted experiments or computations, and are thus not well suited for optimization tasks in chemistry where evaluations of the objectives are often costly (see Sec. 1.2).

Recently, the development of methods for finding the global optimum of non-convex, expensive to evaluate objective functions has gained resurgence as an active field of research. Simplistic approaches consist of random or systematic grid searches. While the advantages of random searches have been demonstrated in the context of hyperparameter optimization for machine learning (ML) models,^{184,505} systematic grid-based approaches like DoE were successfully applied to real-life experiment planning.^{188–190} More sophisticated methods in-

clude genetic algorithms^{506,507} based on evolution strategies,^{206,207} which are for example applied to the optimization of nanoalloy clusters,²⁰⁴ or to resolve electronic spectra of rotamers of organic compounds.²⁰⁵ Yet, such methods still require many evaluations of the objective function.⁵⁰⁸ Bayesian optimization has emerged as a popular and efficient alternative during the last decade (see Sec. 2.3).^{217–220,237,243,248} The typical procedure of Bayesian optimization schemes consists of two major steps. First, an approximation (surrogate) to the merit landscape of the conditions is constructed; and second, a new set of conditions is proposed for the next evaluation based on this surrogate. To that end, Bayesian optimization speculates about the experimental outcome using all previously conducted experiments, and verifies its speculations by requesting the evaluation of a new set of conditions. Several different models have been suggested for approximating the objective function, ranging from random forests (RFs),^{241–243} over Gaussian processes (GPs),^{237,509} to Bayesian neural networks (BNNs).^{246,247} Likewise, a variety of methods for proposing new conditions from the surrogate exists (see Sec. 2.3).^{218,227,237,249,251,252}

Bayesian optimization has been successfully employed for several applications across chemistry. Examples include the property optimization of functional organic molecules,⁵¹⁰ calibrating high-throughput (HT) virtual screening results for organic photovoltaics (OPVs),⁵¹¹ or designing nanostructures for phonon transport.⁵¹² Nevertheless, we identify three challenges in the deployment of Bayesian optimization techniques:

- (i) Domain specificity of the surrogate model: Gaussian processes generally perform well on continuous objective landscapes, while RFs are more suitable for discrete and quasi-discrete landscapes. Unknown experimental response landscapes are more amenable to methods that perform well on a large variety of possible landscapes.
- (ii) Parallelization capabilities: Traditionally, Bayesian optimization methods are sequential, which prevents parallel evaluations of the objective function, for instance, by using multiple experimental platforms or computational resources. For parallel evaluation, however, a procedure enabling the generation of multiple informative parameter points

is needed. Prior works on batched Bayesian optimization,^{238,250,513–515} and non-myopic approaches,⁵¹⁶ aimed to resolve this obstacle, but require additional computation compared to sequential optimization.

- (iii) Computational efficiency: Optimization strategies involving substantial additional computation are only applied efficiently if the time required to suggest a new set of conditions does not significantly exceed the execution time of the experiment. As such, the optimization procedure should be computationally efficient compared to the time required to evaluate the objectives of the considered problem, *e.g.*, to run the experiment.

The *Probabilistic Harvard Optimizer Exploring Non-Intuitive Complex Surfaces* (PHOENICS) algorithm introduced in this study tackles the challenges mentioned above by supplementing ideas from Bayesian optimization with concepts from Bayesian kernel density estimation (BKDE). More technically, we use BNNs (see Chapter 5) to estimate kernel distributions associated with a particular objective function value from observed parameter points. Our approach differs from the traditional use of BNNs in the Bayesian optimization context, where objective function values are predicted from BNNs directly. We construct the approximation to the objective in a simple functional form from estimated kernel distributions. Consequently, the computational cost of PHOENICS scales linearly with the dimensionality of the search space and the number of observations, without the cost of numerous full evaluations of the BNN. We propose an inexpensive acquisition function, which enables intuitive search strategies for efficient batch-wise optimization. This parallelization is achieved by simultaneously proposing multiple parameter points with different sampling policies at negligible additional cost. Those policies are biased towards the exploration or exploitation of the search space tuned by an intuitive hyperparameter. A synergistic effect is observed when proposing batches of parameter points with different sampling policies. Our batching policy not only helps to accelerate the optimization process but also reduces the total number of required function evaluations. It, therefore, constitutes an improvement over

trivial parallelization. In what follows, we start by detailing the mathematical formulation of PHOENICS. We further discuss performance results of PHOENICS on analytic benchmark functions and compare them to other Bayesian optimization methods. Before concluding, we further highlight the applicability of our approach to the *Oregonator*, a model system for chemical reactions, on which we demonstrate the deployment of the proposed optimizer for practical problems in chemistry.

6.2 FORMULATING PHOENICS

For the formulation of PHOENICS, we assume that evaluations of the objective are expensive, where the cost could be related to any budgeted resource such as required execution time, experimental synthesis of chemical compounds, computing resources, and others. Sec. 2.3 details the background on experiment planning as an optimization task. The overall workflow of PHOENICS is schematically represented in Fig. 6.1 and follows the general principles of traditional Bayesian optimization. At each iteration, a surrogate model is constructed, from which new conditions are proposed. In the following, we describe how the surrogate model is constructed (see Fig. 6.1a-c) and how the surrogate model can be biased towards particular sampling strategies, *i.e.*, exploration or exploitation, using an intuitive sampling parameter (see Fig. 6.1d).

6.2.1 APPROXIMATING THE OBJECTIVE FUNCTION

We suggest to use BNNs to estimate the parameter kernel density from the observed parameter points in an autoencoder-like architecture. As such, the BNN is used to non-linearly determine the density of the observed parameter points, \mathbf{x} (see Fig. 6.1b). A particular realization of the BNN represents a map projecting parameter points into the parameter space, *i.e.*, $\text{BNN} : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Thereby, we can construct an estimate to the parameter kernel density, which corresponds to a particular observed objective function value. The use of a BNN for the construction of the surrogate guarantees flexibility in the approximation as it has already been reported that BNNs are versatile function approximators.^{246,247} Details on

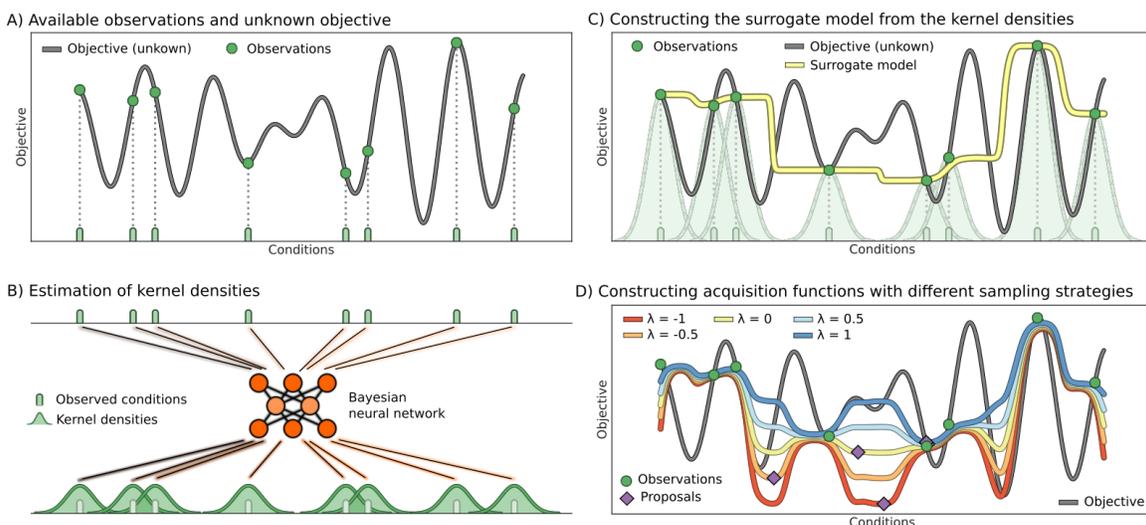


Figure 6.1: Illustration of the workflow of PHOENICS. (A) The unknown, possibly high-dimensional, objective function of an experimental procedure or computation. The objective function has been evaluated at eight different conditions (green), which comprise the set of observations in this illustration. (B) The observed conditions are processed by a Bayesian neural network yielding a probabilistic model for estimating parameter kernel densities. Note, that the probabilistic approach allows for a higher flexibility of our surrogate model compared to standard kernel density estimation. (C) The surrogate model is constructed by weighting the estimated parameter kernel densities with their associated observed objective values. (D) The surrogate can be globally reshaped using a single sampling parameter λ to favor exploration (red) or exploitation (blue) of the parameter space. Reproduced from Ref. [233] with permission from the American Chemical Society.

the BNN architecture are reported in Sec. 6.2.3. The implementation of PHOENICS supports the construction and training of the surrogate model using either the PYMC3 library,^{517,518} or the EDWARD library.⁴⁸² In both cases, the model parameters θ of the BNN are trained *via* variational inference. We start the sampling procedure with 500 samples of burn-in followed by another 1,000 iterations retaining every tenth sample. This protocol was fixed for all tasks.

We can construct an approximation to the kernel density from the distributions of BNN parameters. In particular, for observed conditions \mathcal{D}_n we compute the kernel densities, which are used to approximate the objective function. The kernel density $p_k(\mathbf{x})$ generated from a single observed parameter point, \mathbf{x}_k , (see Fig. 6.1b) can be written in closed form in Eq. 6.1, where $\langle \cdot \rangle$ denotes the average over all sampled BNN architectures, and \mathbf{x}_{pred} denotes the

parameter points sampled from the BNN,

$$p_k(\mathbf{x}) = \left\langle \sqrt{\frac{\tau_n}{2\pi}} \exp \left[-\frac{\tau_n}{2} (\mathbf{x} - \mathbf{x}_{\text{pred}}(\boldsymbol{\theta}; \mathbf{x}_k))^2 \right] \right\rangle_{\text{BNN}}. \quad (6.1)$$

We formulate the approximation to the objective function, α , as an ensemble average of the observed objective function values, f_k , taken over the set of computed kernel densities $p_k(\mathbf{x})$,^{519,520}

$$\alpha(\mathbf{x}) = \frac{\sum_{k=1}^n f_k p_k(\mathbf{x})}{\sum_{k=1}^n p_k(\mathbf{x})}. \quad (6.2)$$

In this ensemble average, each of the constructed distributions $p_k(\mathbf{x})$ is rescaled by the value of the objective function, f_k , observed for the parameter point \mathbf{x}_k (Fig. 6.1c). Note, that α converges to the true objective, f , in the limit of infinitely many distinct function evaluations, as the kernel densities p_k become more peaked due to the increasing precision in the Gaussian prior. Details are provided in Sec. 6.2.3. This approximation to the objective function allows for inexpensive evaluations for any given parameter point \mathbf{x} as repetitive, full BNN evaluations are avoided.

6.2.2 ACQUISITION FUNCTION

In the resulting approximation α we effectively model the expectation value of f for a given parameter point \mathbf{x} based on prior observations. However, the parameter space could contain low density regions, for which the objective function approximation $\alpha(\mathbf{x})$ is inaccurate (see Fig. 6.1c). With these considerations, we propose an acquisition function based on the kernel densities $p_k(\mathbf{x})$ for observations \mathcal{D}_n ,

$$\alpha(\mathbf{x}) = \frac{\sum_{k=1}^n f_k p_k(\mathbf{x}) + \lambda p_{\text{uniform}}(\mathbf{x})}{\sum_{k=1}^n p_k(\mathbf{x}) + p_{\text{uniform}}(\mathbf{x})}. \quad (6.3)$$

The acquisition function differs from the approximation to the objective function (see Eq. 6.2) by an additional term $p_{\text{uniform}}(x)$ in the numerator and the denominator, which denotes the uniform distribution on the domain (see Fig. 6.1d). In the numerator, $p_{\text{uniform}}(x)$ is scaled by a factor λ , referred to as the sampling parameter from hereon. Note, that this modification to α does not affect its convergence behavior (see Sec. 6.2.3 for details). The introduced parameter λ effectively compares the cumulative height of each rescaled density estimate $p_k(\mathbf{x})$ to the uniform distribution. While the $p_k(\mathbf{x})$ are constructed from the knowledge we acquired from previous experiments, p_{uniform} is used as a reference to indicate the lack of knowledge in parameter space regions where little or no information is available yet. The sampling parameter therefore balances between acquired knowledge and the lack of knowledge, which effectively tunes the exploitative and explorative behavior of the algorithm.

Fig. 6.1d illustrates the behavior of PHOENICS on a one-dimensional objective function with different λ values. In this example, the acquisition function is constructed from eight observations indicated in green. Note that the acquisition function approximates the value of the objective function at observed parameter points. Acquisition functions which were constructed from a more positive λ show low values only in the vicinity of the observation with the lowest objective function value. In contrast, acquisition functions that were constructed from a more negative λ show low values far away from any observation. The choice for the value of the exploration parameter λ can, therefore, be directly related to explorative or exploitative behavior when proposing new conditions based on the global minimum of the surrogate. With a large positive value of λ , PHOENICS favors exploitation, while a large negative value favors exploration. When $\lambda = 0$, the acquisition function approximates the objective function itself. Details on the global optimization of the surrogate are provided in the appendix of Ref. [233]. From Fig. 6.1d we see that distinct points in parameter space are proposed based on particular values of the exploration parameter λ . The best choice of λ for a given objective function is *a priori* unknown. However, with the possibility to rapidly construct several acquisition functions with biases towards exploration or exploitation, we can propose multiple parameter points in batches based on different sampling strategies. The

newly proposed parameter points are then evaluated on the black-box optimization function in possibly parallel evaluation runs.

6.2.3 CONVERGENCE OF THE APPROXIMATION TO THE OBJECTIVE FUNCTION

The acquisition introduced in Eq. 6.3 needs to converge to the true, unknown objective function to constitute a suitable approximation on which we can base a Bayesian optimization framework. Here, we argue that this equation indeed presents a surrogate which converges to the objective function in the limit of infinitely many distinct observations. We start our discussion with the construction of the kernel densities p_k as outlined in Eq. 6.1. Note, that $\mathbf{x}_{\text{pred}}(\boldsymbol{\theta}; \mathbf{x}_k)$ are constructed according to Eqs. 6.4 to 6.6, defining the BNN, where \mathbf{x}_k denotes the parameter point associated to the objective function value f_k ,

$$\phi_1 = \tanh(\mathbf{x}_k \cdot \mathbf{w}_0 + \mathbf{b}_0), \quad (6.4)$$

$$\phi_2 = \tanh(\phi_1 \cdot \mathbf{w}_1 + \mathbf{b}_1), \quad (6.5)$$

$$\mathbf{x}_{\text{pred}} = \text{sigmoid}(\phi_2 \cdot \mathbf{w}_2 + \mathbf{b}_2). \quad (6.6)$$

Note, that weights \mathbf{w}_i and biases \mathbf{b}_i can be chosen, such that $\mathbf{x}_{\text{pred}} = \mathbf{x}_k$. In the scenario of a perfectly trained BNN, the kernel densities therefore reduce to

$$p_k(\mathbf{x}) = \sqrt{\frac{\tau_n}{2\pi}} \exp\left[-\frac{\tau_n}{2}(\mathbf{x} - \mathbf{x}_k)^2\right]. \quad (6.7)$$

The precisions, τ_n , of these kernels are sampled from a Gamma distribution, whose expectation value linearly increases with the number of observations, n . We consider the value of the approximation, α , at the position of an observed parameter point, \mathbf{x}_l , in the limit of

infinitely many observations, $n \rightarrow \infty$, where $\tau_n \rightarrow \infty$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \alpha(\mathbf{x}_l) &= \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k p_k(\mathbf{x}_l)}{\sum_{k=1}^n p_k(\mathbf{x}_l)} = \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k \sqrt{\frac{\tau_n}{2\pi}} \exp\left[-\frac{\tau_n}{2}(\mathbf{x}_l - \mathbf{x}_k)^2\right]}{\sum_{k=1}^n \sqrt{\frac{\tau_n}{2\pi}} \exp\left[-\frac{\tau_n}{2}(\mathbf{x}_l - \mathbf{x}_k)^2\right]} \\ &= \lim_{n \rightarrow \infty} \frac{f_l + \sum_{k=1, k \neq l}^n f_k \exp\left[-\frac{\tau_n}{2}(\mathbf{x}_l - \mathbf{x}_k)^2\right]}{1 + \sum_{k=1, k \neq l}^n \exp\left[-\frac{\tau_n}{2}(\mathbf{x}_l - \mathbf{x}_k)^2\right]} = f_l, \end{aligned} \quad (6.8)$$

where we used the fact that the series in the numerator and the denominator are absolute convergent and $\|\mathbf{x}_l - \mathbf{x}_k\| < 1$ for any \mathbf{x}_l and \mathbf{x}_k on the considered domain. When evaluating the approximation α at an arbitrary point \mathbf{x} in the domain, however, we find that the value assumed by the approximation is governed by the observed function value associated with the closest observed parameter point \mathbf{x}_k . We carry out the same limit as in Eq. 6.8,

$$\lim_{n \rightarrow \infty} \alpha(\mathbf{x}) = \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k p_k(\mathbf{x})}{\sum_{k=1}^n p_k(\mathbf{x})} = \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k \sqrt{\frac{\tau_n}{2\pi}} \exp\left[-\frac{\tau_n}{2}(\mathbf{x} - \mathbf{x}_k)^2\right]}{\sum_{k=1}^n \sqrt{\frac{\tau_n}{2\pi}} \exp\left[-\frac{\tau_n}{2}(\mathbf{x} - \mathbf{x}_k)^2\right]} = \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k \beta_k^{\tau_n}}{\sum_{k=1}^n \beta_k^{\tau_n}}, \quad (6.9)$$

where we introduced β_k as $\beta_k(\mathbf{x}) = \exp(-(\mathbf{x} - \mathbf{x}_k)^2/2)$. This expression can be simplified further by dividing both numerator and denominator by the largest β_k , which we denote with $\beta_m = \max_{k=1, \dots, n} (\beta_k)$. Note, that the largest β_k is obtained for the point \mathbf{x}_m in the parameter space, which is the closest to the considered point \mathbf{x} . This simplification leads us to

$$\lim_{n \rightarrow \infty} \alpha(\mathbf{x}) = \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k (\beta_k/\beta_m)^{\tau_n}}{\sum_{k=1}^n (\beta_k/\beta_m)^{\tau_n}} = \lim_{n \rightarrow \infty} \frac{f_m + \sum_{k=1, k \neq m}^n f_k (\beta_k/\beta_m)^{\tau_n}}{1 + \sum_{k=1, k \neq m}^n (\beta_k/\beta_m)^{\tau_n}} = f_m, \quad (6.10)$$

as all $\beta_k/\beta_m < 1$ for $k \neq m$. The approximation is therefore dominated by the objective function value f_m observed for the closest point \mathbf{x}_m in the parameter space. The same

convergence behavior can be observed for the extended expectation calculation,

$$\begin{aligned}
\lim_{n \rightarrow \infty} \alpha(\mathbf{x}_l) &= \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k p_k(\mathbf{x}_l) + \lambda p_{\text{uniform}}(\mathbf{x}_l)}{\sum_{k=1}^n p_k(\mathbf{x}_l) + p_{\text{uniform}}(\mathbf{x}_l)}, \\
&= \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n f_k \sqrt{\frac{\tau_n}{2\pi}} \exp\left[-\frac{\tau_n}{2}(\mathbf{x}_l - \mathbf{x}_k)^2\right] + \lambda p_{\text{uniform}}(\mathbf{x}_l)}{\sum_{k=1}^n \sqrt{\frac{\tau_n}{2\pi}} \exp\left[-\frac{\tau_n}{2}(\mathbf{x}_l - \mathbf{x}_k)^2\right] + p_{\text{uniform}}(\mathbf{x}_l)}, \\
&= \lim_{n \rightarrow \infty} \frac{f_l + \sum_{k=1, k \neq l}^n f_k \exp\left[-\frac{\tau_n}{2}(\mathbf{x}_l - \mathbf{x}_k)^2\right] + \sqrt{\frac{2\pi}{\tau_n}} \lambda p_{\text{uniform}}(\mathbf{x}_l)}{1 + \sum_{k=1, k \neq l}^n \exp\left[-\frac{\tau_n}{2}(\mathbf{x}_l - \mathbf{x}_k)^2\right] + \sqrt{\frac{2\pi}{\tau_n}} p_{\text{uniform}}(\mathbf{x}_l)} = f_l, \quad (6.11)
\end{aligned}$$

where we used the fact that $p_{\text{uniform}}(\mathbf{x}_l)$ is constant and finite. Similar to the previous scenario it can be shown that also the extended approximation approaches the objective function value associated with the closest observed parameter point when evaluated at arbitrary parameter points \mathbf{x} .

6.3 RESULTS AND DISCUSSION

In this section we report the performance of PHOENICS and compare it to four frequently used optimization algorithms: particle swarm optimization (PSO),^{208,209} the covariance matrix adaptation evolution strategy (CMA-ES),^{206,207} and Bayesian optimization based on Gaussian processes and RFs. PSO is implemented in the PYSWARMS python module.⁵²¹ An implementation of CMA-ES is available in the CMA module.⁵²² Default settings as provided by the modules have been used for both optimization algorithms. The SPEARMINT package performs Bayesian optimization using Gaussian processes and the predictive entropy acquisition function.^{237,509} Batch optimization is implemented in SPEARMINT *via* estimating the expected objective value for future evaluations. The SMAC package employs RF models and allows for batch optimization by running multiple RF instances sharing the same set of samples.²⁴¹⁻²⁴³

Chemical systems can have complex, qualitatively different response surfaces. As chem-

ical reactions are time-consuming to evaluate, we assess the performance of each of these three algorithms on a set of 15 synthetic benchmark functions covering a large range of qualitatively diverse response surfaces for problems in chemistry. The employed functions are well-established benchmarks and include continuous and convex, non-convex, and discrete functions with possibly multiple global minima. A complete list of the objective functions and their global minima is provided in the appendix of Ref. [233]. For reliable performance estimates, we executed 20 independent optimization runs initialized with different random seeds, unless noted otherwise. During each run, we record the lowest achieved objective function value after each iteration. We compare the average lowest achieved objective function values by relating to results from simple random searches. Each random search was run for 10^4 objective function evaluations, and results were averaged over 50 independent runs initialized with different random seeds. The average lowest achieved objective function values of the random search runs are summarized in the appendix of Ref. [233]. Our benchmark calculations indicate that Bayesian-based optimization algorithms outperform PSO and CMA-ES. Specifically, after 200 function evaluations, PSO and CMA-ES yield significantly higher deviations from the global optimum than the three studied Bayesian optimization algorithms. Even when increasing the number of function evaluations by an order of magnitude, from 200 to 2,000, PSO (and CMA-ES) fail to achieve lower deviations than the ones obtained with Phoenix after 200 evaluations for 12 (and 13) out of 15 objective functions. We therefore restrict our further analyses and comparisons to only PHOENIX, GP optimization and RF optimization. The benchmark results conducted with PSO and CMA-ES are detailed in the appendix of Ref. [233].

6.3.1 ANALYTIC BENCHMARKS

PHOENIX was set up with three different values for the sampling parameter, $\lambda \in \{-1, 0, 1\}$, to assess the effectiveness of a particular parameter choice. In Fig. 6.2 we report the number of objective function evaluations required by each of the optimization algorithms to reach objective function values lower than the average lowest value found in random searches. Op-

timization traces for these runs on all 15 objective functions are reported in the appendix of Ref. [233]. We find that GP optimization, as implemented in SPEARMINT, quickly finds the global minimum if the objective function is strictly convex. In contrast, RF optimization, as implemented in SMAC, quickly finds the global minimum of objective functions with a discrete co-domain. The performance of PHOENICS varies with different values of the sampling parameter λ . When favoring exploitation over exploration, *i.e.*, $\lambda > 0$ the algorithm performs better if the objective function features narrow and well defined funnels (*e.g.*, *Ackley* in Fig. 6.2a or *Schwefel* in Fig. 6.2c). With this choice for the sampling parameter, the algorithm is slightly biased towards exploring the local region around the current optimum. This behavior, however, is unfavorable in other cases, for instance, when the objective function has a discrete co-domain (*e.g.*, *dAckley* in Fig. 6.2d). Since parameter points in the vicinity to the current optimum likely yield the same value if the objective function is discrete or quasi-discrete, PHOENICS performs better on such objective functions when favoring exploration over exploitation, *i.e.*, $\lambda < 0$.

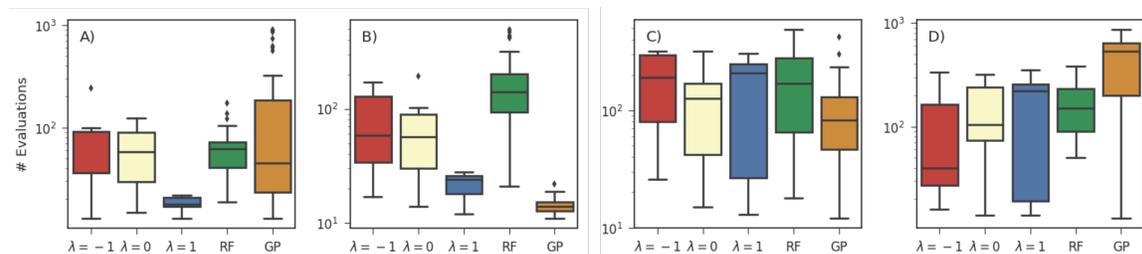


Figure 6.2: Number of objective function evaluations required to reach objective function values lower than the average lowest achieved values of random searches with 10^4 evaluations for PHOENICS ($\lambda \in \{-1, 0, 1\}$), RFs and Gaussian processes. Results are reported for the Ackley (A), Dejong (B), Schwefel (C) and dAckley (D) objective functions. Details on the benchmark functions are provided in the appendix of Ref. [233]. Reproduced from Ref. [233] with permission from the American Chemical Society.

6.3.2 DEVELOPING A COLLECTIVE SAMPLING STRATEGY

The dependence of the performance of PHOENICS on the sampling parameter λ could be eliminated by marginalizing over this parameter. Marginalization over the sampling parameter would effectively average out the advantageous effects of a bias towards exploitation for some objective functions and towards exploration for other objective functions. The shape

of the objective function is *a priori* unknown, such that suitable choices of the sampling parameter cannot be determined beforehand. However, since the sampling parameter can be directly related to the explorative and exploitative behavior of the algorithm, we follow an approach to taking full advantage of the sampling policy. We suggest proposing parameter points based on multiple different sampling parameter values. Note that values of λ are chosen beforehand and are kept fixed throughout the optimization procedure.

Given a set of observations \mathcal{D}_n the construction of several objective surrogates with different values of λ is computationally inexpensive. This feature allows us to suggest multiple parameter points at each optimization iteration, which are proposed from more explorative and more exploitative parameter values, at almost no additional cost. With the observations on the simple benchmarks, we would expect a synergistic effect of this batch optimization over sequential optimization with a single sampling parameter value. As parameter points can be proposed with both biases towards exploration and exploitation, we expect the number of required objective function evaluations to decrease. Suggesting a batch of parameter points in one optimization step also allows for the parallel evaluation of all proposed points, which accelerates the optimization process. The behavior of the three studied optimization algorithms under parallel optimization on the *Ackley* objective function is highlighted in Fig. 6.3a. Full results on the entire benchmark set are reported in the appendix of Ref. [233]. Fig. 6.3a depicts the minimum achieved objective function values for different runs with a different number of parallel evaluations of the objective function averaged over 20 independent runs. The minimum achieved objective function values are presented per objective function evaluations (left panels) and per batch evaluations (right panels).

We find that both SPEARMINT and SMAC achieve low objective function values in fewer batches with an increasing number of points, p , proposed in each batch. While increasing the number of samples per batch initially improves the performance significantly with respect to the number of proposed batches, this advantageous effect quickly levels off until there is no further improvement beyond six samples per batch. However, when comparing the minimum achieved objective function values with respect to the total number of objective

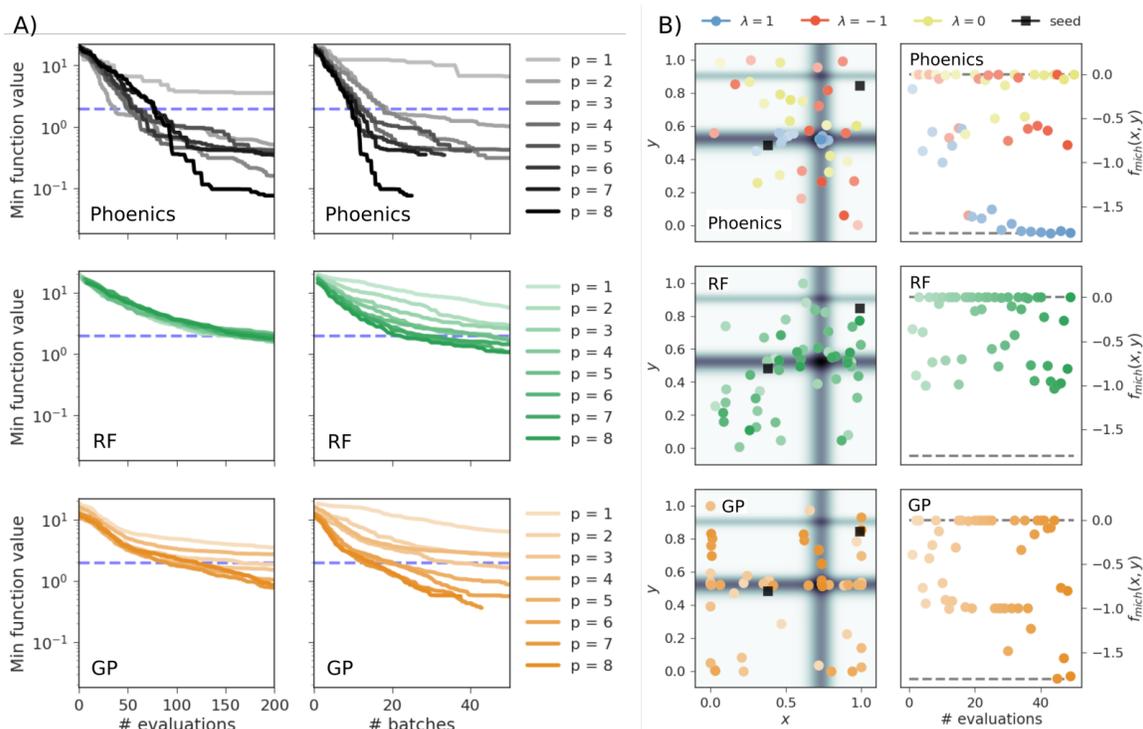


Figure 6.3: (A) Average minimum objective function values for the Ackley function achieved in 20 independent runs of the three optimization algorithms studied in this work: PHOENICS, spearmint (GP) and SMAC (RF). For each run a different number of proposed samples p was evaluated in parallel. Minimum achieved objective function values are reported with respect to the total number of objective function evaluations and the number of evaluated batches. The dashed blue lines denote the minimum achieved error after 10^4 of random search for reference. (B) Progress of sample optimization runs of the three studied optimization algorithms on the two dimensional Michalewicz function. PHOENICS proposed a total of three samples per batch, which were then evaluated in parallel. Each sample was suggested based on a particular value of the exploration parameter $\lambda \in \{-1, 0, 1\}$. Left panels illustrate the parameter points proposed at each optimization iteration while right panels depict the achieved objective function values. Depicted points are more transparent at the beginning of the optimization and more opaque towards the end. Starting points for the optimization runs are drawn as black squares. Reproduced from Ref. [233] with permission from the American Chemical Society.

function evaluations, we did not observe any significant difference between runs with a different number of samples per batch. In contrast, PHOENICS shows different behavior. Our algorithm does not only reach lower objective function values with fewer batches when proposing more samples per batch but also shows better performance when considering the total number of function evaluations. This synergistic effect demonstrates that PHOENICS indeed benefits from proposing points with multiple sampling strategies in cases in which proposed samples are evaluated sequentially. The performance improvement of PHOENICS when proposing parameter points in batches at each optimization iteration is demonstrated

on all 15 considered objective functions in the appendix of Ref. [233]. We ran our optimizer with four different sampling strategies using sampling parameter values evenly spaced across the $[-1, 1]$ interval. All four proposed parameter points are then evaluated before we started another optimization iteration. For this particular batching protocol, we find that PHOENICS outperforms RF based optimization on all benchmark functions and GP based optimization on 12 out of 15 benchmark functions. *If and only if* the objective function is convex, GP optimization finds lower objective function values. We observe the aforementioned synergistic effect of batch optimization for 12 out of 15 benchmark functions. Despite reducing the number of iterations by a factor of 4, the achieved objective function values were found to be lower than values achieved in sequential optimizations with all three considered fixed sample parameter values.

We suggest that this improved performance of the algorithm is due to the trade-off between exploration and exploitation. The exploration samples systematically sample the parameter space and ensure that the algorithm does not get stuck in local minima, while the exploitation samples explore the local environment of the current global minimum. This sampling behavior is illustrated in Fig. 6.3b for the *Michalewicz* function. The optimization runs on the Michalewicz function were all started from the same two random samples illustrated in black for all three investigated optimization algorithms. Bayesian optimization based on Gaussian processes, as implemented in SPEARMINT (lower panels), tends to sample many parameter points close to the boundaries of the domain space in this particular example. RF optimization, as implemented in SMAC (central panels), however, shows a higher tendency of exploring the parameter space. PHOENICS (upper panels) starts exploring the space and quickly locates a local minimum in the vicinity of one of the initial samples. After finding this local minimum, samples which are proposed based on a more exploitative (positive) value of the sampling parameter λ explore the local environment of this local minimum while samples proposed from more explorative (negative) values of λ explore the entire parameter space. As soon as the exploration points find a point in the parameter space with a lower value of the objective function, the exploitation points jump to this new region and

locally explore the vicinity of the current best sample.

Overall we have demonstrated that the value of the sampling parameter λ in the proposed acquisition function influences the behavior of the optimization procedure towards a more explorative behavior for more negative values of this parameter and a more exploitative behavior for more positive parameter values. Batched optimization improves the performance of PHOENICS, even in terms of total objective function evaluations, and reduces the number of required optimization iterations.

6.3.3 INCREASING THE NUMBER OF DIMENSIONS

Practical chemistry problems are typically concerned with more than just two parameters (see Sec. 2.3). Environmental conditions and experimental device settings can influence chemical reactions and computational studies frequently employ parameters to describe the system of interest. In this section, we illustrate the performance of PHOENICS in parameter spaces with dimensions $k > 2$. We evaluate the performance of the three optimization algorithms on the same objective function subset, but now successively increase the dimensionality of the parameter space from two to 20. Based on the results on batch optimization, we ran GP and RF optimization with one point per batch and the optimization algorithm introduced in this chapter with four points per batch on each considered benchmark function. Exploration parameter values were chosen to be evenly spaced across the $[-1, 1]$ interval. For better comparisons, we report the average deviation of the lowest encountered objective function value from the global minimum of each function taken over 20 independent optimization runs. Average deviations achieved by each of the optimization algorithms after 200 objective function evaluations are depicted in Fig. 6.4.

We observe that PHOENICS maintains its rapid optimization properties for a variety of different objective functions, even when increasing the number of dimensions. In the case of the *Ackley* function (Fig. 6.4a) PHOENICS appears to find and explore the major funnel close to the global optimum faster than the other two optimization algorithms regardless of the number of dimensions. The paraboloid (Fig. 6.4b) is an easy case for the GP in low

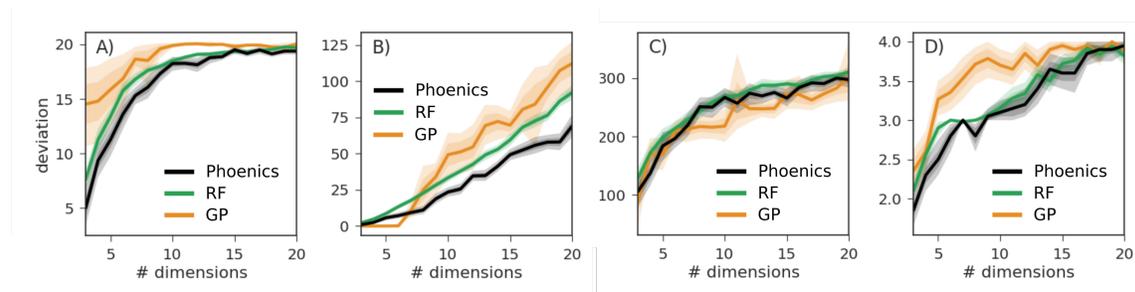


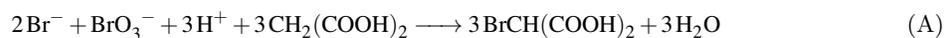
Figure 6.4: Average deviations taken over 20 independent runs between the lowest encountered objective function value and the global minimum achieved after 200 objective function evaluations for different parameter set dimensions. Results are reported for *Ackley* (A), *Dejong* (B), *Schwefel* (C) and *dAckley* (D). Uncertainty bands illustrate bootstrapped estimates of the deviation of the means with one and two standard deviations. Reproduced from Ref. [233] with permission from the American Chemical Society.

dimensions, but is optimized the fastest by PHOENICS when considering parameter spaces with seven or more dimensions. No major differences are observed for the *Schwefel* function (Fig. 6.4c). However, in the case of a discrete objective function (Fig. 6.4d) PHOENICS seems to have a slight advantage over the other two optimizers for lower dimensions and performs about as well as RF optimization for higher dimensions.

6.4 APPLICATIONS TO CHEMISTRY

In this section, we demonstrate the applicability of PHOENICS on the *Oregonator*, a model system of a chemical reaction described by a set of non-linear coupled differential equations.⁵²³ In particular, we demonstrate how PHOENICS can be employed to propose a set of conditions for an experimental procedure. The experiment can then be executed with the proposed conditions, and the results of the procedure are reported back to PHOENICS. With this feedback, PHOENICS can make more informed decisions and, thus, provides more promising experimental conditions, to eventually discover the optimal set of conditions. Most chemical reactions lead to a steady-state, *i.e.*, a state in which the concentrations of involved compounds are constant in time. While chemical reactions described by linear differential equations always feature such a steady-state, more complex dynamics phenomena can arise for reactions described by sets of non-linear coupled differential equations. With the right choice of parameters, such differential equations may have a stable limit cycle,

leading to periodic oscillations in the concentrations of involved compounds.^{524,525} One of the earliest discovered reactions featuring a stable limit cycle is the Belousov-Zhabotinsky reaction.^{526,527} This network of chemical reactions involves temporal oscillations of $[\text{Ce}^{\text{IV}}]$ and $[\text{Ce}^{\text{III}}]$. The entire reaction network can be written as a set of three subreactions listed in reactions A, B and C. For details on the mechanism, we refer to a summary in the literature.^{523,525,527–529}



Models at different levels of complexity have been developed to describe the dynamical behavior of the Belousov-Zhabotinsky reaction.^{524,529–531} One of the simplest models of this reaction is the Oregonator.⁵²³ The Oregonator consists of a set of three coupled first order non-linear differential equations for three model compounds X , Y and Z , which are shown in Eqs. 6.12 to 6.14. These equations involve five reaction constants k_i , a stoichiometric factor f , determined by the prevalence of one subreaction over another subreaction, and the concentration of two additional chemical compounds A and B . A map of Eqs. 6.12 to 6.14 to reactions A, B and C is outlined in Ref.⁵²³

$$\frac{dX}{dt} = k_1AY - k_2XY + k_3BX - 2k_4X^2, \quad (6.12)$$

$$\frac{dY}{dt} = -k_1AY - k_2XY + fk_5Z, \quad (6.13)$$

$$\frac{dZ}{dt} = k_3BX - k_5Z. \quad (6.14)$$

The set of differential equations in the Oregonator can be reduced into a dimensionless form, such that the number of correlated parameters is reduced to a smaller set of independent parameters. This reduced version of the Oregonator includes three dimensionless variables α , η and ρ , which describe the concentration of chemical species, and four dimensionless

reaction constants. PHOENICS is used to reverse engineer the set of reaction conditions consisting of three initial concentrations (α_0 , η_0 and ρ_0) and four reaction constants (q , s , w and f) from the concentration traces computed in the original publication.⁵²³ Parameter values yielding the target concentration traces are reported in Tab. 6.1 The goal is twofold: (i) finding a set of conditions for which the dynamical behavior qualitatively agrees with the behavior of the target, *i.e.*, finding chemical oscillations and (ii) fine-tuning these conditions such that we reproduce the dynamical behavior on a quantitative level. We aim to achieve these two goals while keeping the number of function evaluations, *i.e.*, the number of experiments to run, to a minimum. Note, that this constraint along with the dimensionality of the parameter space implies that grid searches or gradient-based algorithms are not suited for this problem.

Table 6.1: Reaction parameters of the reduced Oregonator model for the Belousov-Zhabotinsky reaction. Target parameters induce the existence of a limit cycle, from which chemical oscillations emerge. For finding these target parameters via optimization we constrained the domain space to the reported ranges. All reported quantities are dimensionless.

Parameter	Target value	Range
s	77.27	0...100
w	0.1610	0...1
q	$8.375 \cdot 10^{-6}$	$10^{-8} \dots 10^{-4}$
f	1	0...5
α_0	$2.0 \cdot 10^7$	$10^4 \dots 10^9$
η_0	$3.3 \cdot 10^3$	$10^3 \dots 10^5$
ρ_0	$4.1 \cdot 10^4$	$10^3 \dots 10^6$

PHOENICS was run in parallel proposing four samples per batch with λ equally spaced on the $[-1, 1]$ interval. We compare the performance to PSO in the PYSWARMS module, CMA-ES in the CMA module, GP optimization in SPEARMINT and RF optimization in SMAC. Each of the five optimization algorithms was used in 50 independent optimization runs for 150 evaluations. All optimization procedures were carried out on a constrained parameter space reported in Tab. 6.1. Note that different choices of parameter sets within this bounded domain can result in quantitatively and qualitatively different dynamical behavior. In particular, parameter choices close to the target result in oscillatory behavior, for which

the reduced concentrations α , ρ and η change periodically over time, while other parameter choices can break the limit cycle and create a stable fixed point instead.^{523,528,531}

Concentration traces for a sampled set of reaction parameters were computed with a fourth-order Runge-Kutta integrator with adaptive time stepping. The integrator was run for a total of 10^7 integration steps covering 12 full concentration oscillations for the target parameter set. Sampled concentration traces are compared to the target concentration traces after a cubic spline interpolation. The distance (*loss*) between the sampled traces and the target traces is calculated as the Euclidean distance between the points in time at which a concentration trace reaches a dimensionless concentration value of 100. Average achieved losses for all three optimization algorithms are displayed in Fig. 6.5. Loss values between 300 and 500 indicate that the periodicity of the predicted concentration traces resembles the periodicity of the target traces, *i.e.*, the predicted traces qualitatively agree with the target. A quantitative agreement, *i.e.*, matching traces, is only achieved for loss values lower than about 100.

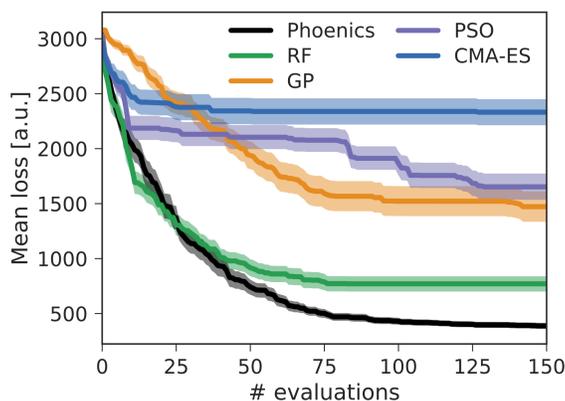


Figure 6.5: Average achieved losses for finding reaction parameters of the reduced Oregonator model achieved by the five optimization algorithms employed in this study. Correct periodicities of the concentration traces are achieved for losses lower than 500. Uncertainty bands illustrate bootstrapped deviations on the mean for one standard deviation. Reproduced from Ref. [233] with permission from the American Chemical Society.

Fig. 6.6 shows concentration traces associated with the lowest loss achieved by each of the three optimization algorithms across all 50 independent runs. PHOENICS is the only algorithm reproducing the target dynamic behavior qualitatively and quantitatively within

150 optimization iterations. RF optimization only finds parameter sets which qualitatively agree with the target. GP optimization finds only in rare occasions concentration traces in qualitative agreement with the target. Both PSO and CMA-ES consistently yield high losses for the first 75 evaluations, after which PSO slightly improves but never reaches the degree of agreement achieved by PHOENICS.

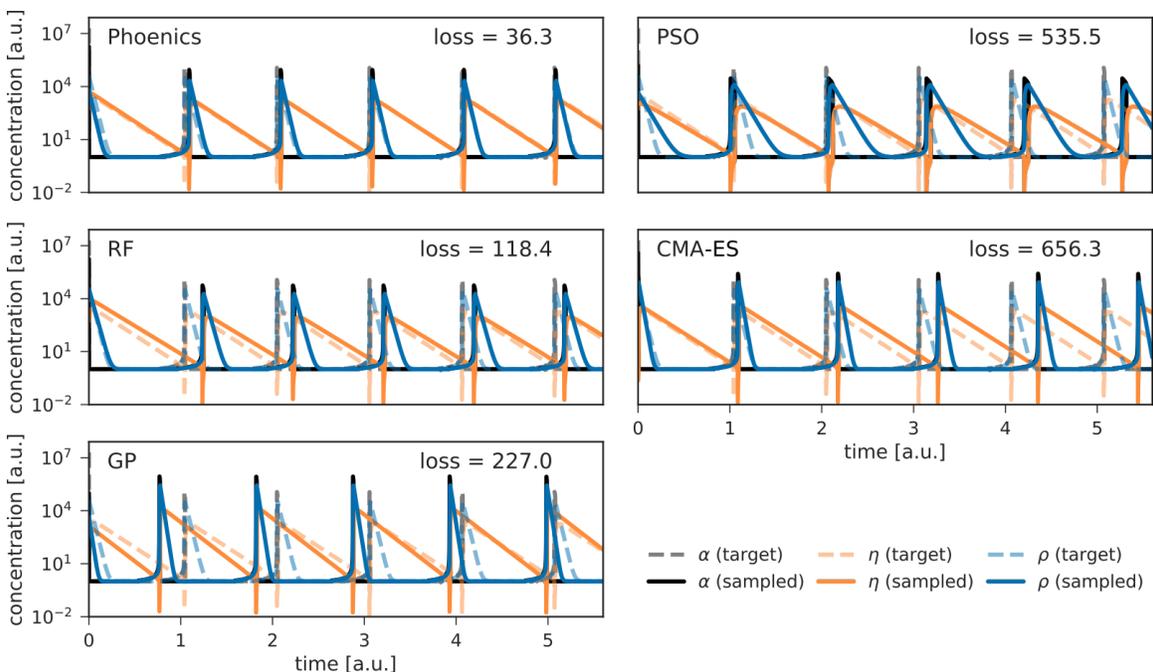


Figure 6.6: Time traces of dimensionless concentrations of compounds in the *Oregonator* model. Target traces are depicted with solid, transparent lines while predicted traces are shown in dashed, opaque lines. Traces were simulated for a total of 12 dimensionless time units, but are only shown for the first six time units for clarity. Reproduced from Ref. [233] with permission from the American Chemical Society.

6.5 CONCLUSION AND OUTLOOK

We introduced PHOENICS, an algorithm for global optimization in the context of chemistry and experimentation. PHOENICS is designed for scenarios where the merit of a set of conditions is evaluated *via* experimentation or expensive computations, which can possibly be parallelized. Our probabilistic optimizer combines Bayesian optimization with conceptual aspects of BKDE. As such, our algorithm is well suited for applications where evaluations of the objective function are expensive with respect to budgeted resources such as time

or money. Through an exhaustive benchmark study, we showed that PHOENICS improves over optimization strategies based on particle swarms or evolutionary approaches, as well as on existing Bayesian global optimization methods and avoids redundant evaluations of the objective. We formulate an inexpensive acquisition function balancing the explorative and exploitative behavior of the algorithm. This acquisition function enables intuitive sampling policies for an efficient parallel search of global minima. By leveraging synergistic effects from running multiple sampling policies in batches, the performance of the algorithm improves and requires a reduced total number of objective function evaluations. The applicability of PHOENICS was highlighted on the *Oregonator*, a model system describing a complex chemical reaction network. PHOENICS was able to determine the set of seven experimental conditions reproducing a target dynamic behavior in the concentrations of involved chemical species. High degrees of qualitative and quantitative agreement could be achieved with only 100 merit evaluations of proposed conditions despite the rich solution space containing both steady-state systems and chemical oscillators. We believe that PHOENICS has the potential to be applied to a wide range of tasks, from the optimization of reaction conditions and material properties, over control of robotics systems, to circuit design for quantum computing.^{532,533} All in all, we recommend PHOENICS for efficient optimizations of scalar, possibly non-convex, black-box unknown objective functions.

7

GRYFFIN: An algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications to chemistry

Apart from minor modifications, this chapter originally appeared as:²³⁴

Gryffin: An algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications in chemistry. Florian Häse, Loïc M. Roch and Alán Aspuru-Guzik. *arXiv preprint* arXiv:2003.12127 (2020).

ABSTRACT

Designing functional molecules and advanced materials requires complex interdependent design choices: tuning continuous process parameters such as temperatures or flow rates, while simultaneously selecting categorical variables including catalysts or solvents. To date, the development of data-driven experiment planning strategies for autonomous workflows has primarily focused on continuous process parameters. Yet, efficient strategies for the selection of categorical variables are required to further accelerate scientific discovery. In this chapter, we introduce GRYFFIN as a general-purpose framework for the autonomous selection of optimal categorical variables driven by expert knowledge. GRYFFIN augments Bayesian optimization with kernel density estimation using smooth approximations to categorical distributions. By leveraging domain knowledge from physicochemical descriptors, GRYFFIN can significantly accelerate the search for promising molecules and materials. GRYFFIN can

further highlight relevant descriptors to inspire physical insights. In addition to comprehensive benchmarks, we demonstrate the capabilities and performance of GRYFFIN on three examples in materials science and chemistry: (i) the discovery of non-fullerene acceptors for organic solar cells, (ii) the design of hybrid organic-inorganic perovskites for light-harvesting, and (iii) the identification of ligands and process parameters for Suzuki-Miyaura reactions. Our observations suggest that GRYFFIN, in its simplest form without descriptors, constitutes a competitive categorical optimizer compared to state-of-the-art approaches. However, when leveraging domain knowledge provided *via* descriptors, GRYFFIN can optimize at considerably higher rates and refine this domain knowledge to spark scientific understanding.

7.1 CATEGORICAL DESIGN CHOICES IN MATERIALS SCIENCE AND CHEMISTRY

The discovery of functional molecules and advanced materials is recognized as one of the fundamental obstacles to the development of emerging and future technologies to face immediate challenges in clean energy, sustainability, and global health (see Chapter 1).^{7,534} To date, accelerations of scientific discovery workflows across chemistry, materials science, and biology have largely been driven by combinatorial high-throughput (HT) strategies with automated experimentation equipment.^{180–182,535,536} Despite remarkable successes with HT approaches (see Chapter 9),^{193,537–540} the combinatorial explosion of molecular and materials candidates renders exhaustive evaluations on large scales impossible. This limitation can be alleviated by adaptive search strategies which selectively explore the search space and only evaluate the most promising materials candidates.⁵⁴¹ Autonomous platforms have been suggested as a next-generation approach to experimentation for accelerated scientific discovery.^{81,82,84,417} These platforms augment automated experimentation systems with data-driven algorithmic strategies to continuously plan new experiments inspired by previously collected measurements.

Recently, data-driven experiment planning has experienced increased attention across various applications including the search for antimicrobial peptides,⁵⁴² the synthesis of organic molecules,^{508,543} the discovery and crystallization of polyoxometalates,⁵⁰⁴ the discovery of

metallic glasses,⁵⁴⁴ the optimization of carbon dioxide-assisted nanoparticle deposition,⁵⁴⁵ and the creation of Bose-Einstein condensates.⁵⁴⁶ Motivated by the successes of data-driven experiment planning, the development, deployment, and benefits of autonomous workflows for scientific discovery over conventional experimentation strategies are being actively explored.⁸³ For example, autonomous platforms have been reported for the optimization of reaction conditions in the context of flow chemistry,^{503,547,548} the unsupervised growth of carbon nanotubes,^{203,549} autonomous synchrotron X-ray characterization,^{550,551} the discovery of thin-film materials (see Chapter 11),⁹⁰ the synthesis of inorganic photoluminescent quantum dots,⁵⁵² and the discovery of photostable quaternary polymer blends for organic photovoltaic (OPV) (see Chapter 12).⁹¹

Although autonomous platforms appear to be on the rise, and data-driven approaches are emerging as viable experiment planning strategies, the examples mentioned earlier mostly targeted experimentation tasks with continuous process parameters. However, scientific discovery in chemistry and materials science typically involves the simultaneous optimization of continuous and categorical variables, such as the selection of a catalyst or solvent, which cannot be targeted efficiently with continuous methods. Approaches to the data-driven selection of categorical parameters are often handcrafted and involve human decisions, which can adversely affect the experimentation throughput. Examples of experimentation workflows involving the selection of categorical variables with partial human interaction have been demonstrated in the context of reaction optimization for flow chemistry.^{191,553,554} The lack of a general-purpose approach to the data-driven optimization of categorical variables is a challenge to autonomous discovery workflows and appears as a significant obstacle to the wide-spread adoption of autonomous experimentation platforms.

The machine learning (ML) community is actively exploring algorithmic approaches to the data-driven selection of categorical variables for hyperparameter optimization,²³⁷ and control parameters in robotics.⁵⁵⁵ However, these applications are different from chemistry and materials science, where categorical variables can usually be characterized by a notion of similarity between individual choices. For example, co-polymers for hydrogen produc-

tion can be synthesized from categorical monomers which differ in their reactivity.⁵⁵⁶ More generally, similarity measures between molecules and materials can be introduced based on their physical, chemical, and structural properties. An experiment planning strategy which actively leverages physicochemical descriptors of candidate materials would be most desirable to (i) accelerate scientific discovery and (ii) gain scientific insights to inspire the design of even more promising molecules and materials that are not included in the search library.

In this chapter, we introduce GRYFFIN, a global optimization strategy for the selection of categorical variables in autonomous workflows. GRYFFIN implements an efficient Bayesian optimization framework leveraging kernel density estimation directly on the categorical space, which can be accelerated with domain knowledge in the form of physicochemical descriptors by locally redefining the metric on the categorical space. GRYFFIN can construct more informative descriptors on-the-fly to highlight the relevance of the provided descriptors and inspire scientific interpretations while identifying desired categorical options at a faster rate. We highlight the applicability and performance of GRYFFIN on a set of synthetic benchmark functions and three real-world tasks: the discovery of small molecule non-fullerene acceptors for organic solar cells (OSCs), the discovery of hybrid organic-inorganic perovskites (HOIPs) for light-harvesting and the combined selection of ligands and process parameters for Suzuki-Miyaura coupling reactions. We identify three key advantages of GRYFFIN over state-of-the-art approaches to categorical optimization: (i) it provides a competitive framework for the descriptor-less optimization of categorical variables, (ii) can optimize at significantly higher rates with provided and refined descriptors which can inspire scientific insights, and (iii) integrates with continuous optimization strategies to enable the robust and efficient optimization of mixed continuous-categorical domains for sequential and batched workflows.

7.2 BACKGROUND AND RELATED WORK

Experiment planning can be formulated as an optimization task, where we consider a set of controllable parameters within a defined domain, $\mathbf{z} \in \mathcal{Z}^n$, and an experimental response,

$f(z)$, for each of the parameter choices within the domain. In the context of reaction optimization, the controllable parameter could, for example, include the reaction temperature, the amount of solvent, and the choice of the catalyst. At the same time, the experimental response could be quantified via the rate at which the desired product is generated. The optimization domain is also referred to as the *design space* or the *candidate space*. The optimization task in experiment planning consists in the identification of specific parameter values, $\mathbf{z}^* \in \mathcal{Z}^n$, which yield the desired experimental outcome, $f(\mathbf{z}^*)$. For simplicity, we will consider minimization tasks from hereon, *i.e.*, we formulate f such that $\mathbf{z}^* = \underset{\mathbf{z} \in \mathcal{Z}}{\operatorname{argmin}} f(\mathbf{z})$ corresponds to the desired experimental result. The optimization task can be approached with a closed-loop strategy, which iteratively evaluates a set of options \mathbf{z}_j and records to associated responses, $f_j = f(\mathbf{z}_j)$, to gradually collect a set of observations, $\mathcal{D}_n = \{\mathbf{z}_j, f_j\}_{j=1}^n$, as feedback to the experiment planning strategy. Sec. 2.3 discusses the interpretation of scientific discovery as an optimization task if the controlled parameters are continuous. Here, we focus on categorical choices, which pose additional challenges due to the lack of a natural ordering between individual parameter values.

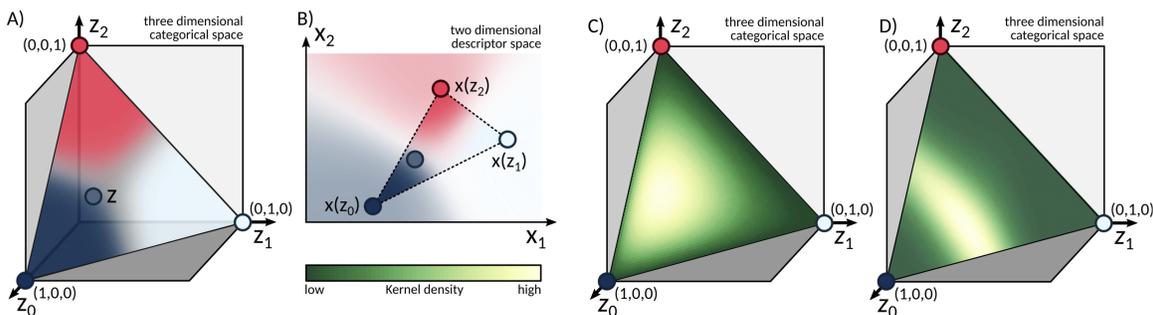


Figure 7.1: Illustrations of the naïve and the static GRYFFIN strategies for the (descriptor-guided) optimization of categorical variables. (A) Illustration of a categorical variable with three options represented on a simplex. Color contours indicate the affiliation of any given point on the simplex to one of the categorical options at the corners. (B) Representation of a continuous descriptor space, where descriptors are associated with the categorical options shown in panel A. Color contours indicate the affiliation with individual options of the categorical variable. (C) Illustration of a generic kernel density on the simplex modeled with a concrete distribution. (D) Illustration of a descriptor-guided transformation of the kernel density shown in panel C based on the descriptors shown in panel D. Reproduced with permission from Ref. [234].

Regardless of the optimization strategy, the confidence of having identified the best performing candidate in the search space increases with the number of evaluated candidates.

In the best case, only one evaluation is required, while in the worst case, all candidates in the search space need to be evaluated. The choice of the optimization strategy modulates the chance of having identified the best performing candidate after a certain number of evaluations and, consequently, the average fraction of the candidate space that needs to be evaluated to identify the most desired one. Straightforward search strategies rely on exhaustive random^{184–186} or systematic^{188–190} evaluations of all candidates without leveraging any feedback from collected responses to refine the search policy. In the absence of accurate prior expectations on the performance of individual candidates, both random and systematic search strategies require the evaluation of 50 % of all candidates, on average, to identify the best performing candidate and are therefore only applicable to relatively small search spaces. Nevertheless, exhaustive strategies are massively parallelizable and thus well suited for high-throughput experimentation. Genetic algorithms and evolutionary strategies^{199–201} extend the idea of a random exploration of the search space but condition their exploration policies on a population of candidate solutions which have already been evaluated. In contrast to a globally random exploration, new candidates are selected based on local perturbations on the population of the best performing candidates. During the optimization, the better performing candidates substitute the poorly performing candidates.²⁰²

Bayesian optimization^{216,217} has recently gained increased attention as a competitive global optimization strategy across various fields,^{221,222} including automatic ML,^{223–225} and experimental design.^{228,229,231} Extensions of Bayesian optimization frameworks to categorical parameter domains are under active development. One approach consists in the representation of categorical parameters as one-hot encoded vectors.^{237,557,558} This representation expresses the j -th option of a categorical variable \mathbf{z} with n different options, $\mathbf{z} = \{z_1, \dots, z_n\}$, as an n -dimensional vector with elements $z_i = \delta_{ij}$ for $1 \leq i \leq n$, which can be interpreted as the corners of an n -dimensional simplex, $\mathbf{z} \in \Delta^{n-1} = \{\mathbf{z} \in \mathbb{R}^n | z_i \in [0, 1] \text{ and } \sum_{i=1}^n z_i = 1\}$ (see Fig. 7.1a). Standard Bayesian optimization strategies for continuous parameter domains can be deployed on these one-hot encoded categorical variables, such that even optimizations of mixed continuous-categorical domains are possible. The choices for future evaluations are

determined by projecting promising candidates from the continuous space to the one-hot boundaries. This strategy presents two limitations: (i) redundancies in the projection arise from the fact that the continuous optimization space contains an additional degree of freedom compared to the categorical domain, and (ii) the one-hot encoding imposes an equal measure of covariance between all choices of the categorical variables such that we cannot account for imbalanced similarities. Redundancies in the projection can be reduced by imposing constraints on the acquisition function on the continuous domain. For example, the acquisition function can be modified such that covariances are computed after the projection operation.^{559,560} This modification results in a stepwise defined acquisition function from which choices for future evaluations can be suggested directly. However, stepwise functions are generally more challenging to optimize than smooth functions, and this modification still retains equal covariance measures between individual choices of categorical variables.

7.3 FORMULATING GRYFFIN

Building upon previous works (see Sec. 7.2), we base the formulation of GRYFFIN on a one-hot encoding of categorical variables. Instead of constructing the surrogate on the continuous space spanned by the one-hot encoded categorical choices, GRYFFIN aims to support the surrogate on the simplex to avoid projection redundancies. To this end, we model categorical parameters as random variables and extend the recently reported PHOENICS approach²³³ from continuous domains to categorical domains. The surrogate is constructed from reweighted kernel density estimates for the categorical parameters, similar to the approach outlined in Chapter 6. Beyond the implementation of kernel density based Bayesian optimization on categorical domains, we further demonstrate how physical and chemical domain knowledge can be used to transform the surrogate to accelerate the search and how this bias can be refined during the optimization to identify and interpret relevant domain knowledge.

7.3.1 CATEGORICAL OPTIMIZATION WITH NAÏVE GRYFFIN

Naïve GRYFFIN constructs kernel densities by extending the one-hot encoding of categorical options to the entire simplex, *i.e.*, we consider $\mathbf{z} \in \Delta^{n-1}$. The largest entry of any given point \mathbf{z} can be used to associate this point to a realizable option ζ (see Fig. 7.1a). Various probability distributions with support on the simplex have been introduced in the past. The Dirichlet distribution, for example, constitutes the conjugate prior to the categorical distribution.⁵⁶¹ Another example is the logistic normal distribution which ensures that the logit of generated samples follows a standard normal distribution.⁵⁶² While both of these distributions are widely used, their deployment in a computational graph is numerically involved due to demanding inference and sampling steps. Directed probabilistic models can be implemented at low computational cost if stochastic nodes of such graphs can be reparameterized into deterministic functions of their parameters and stationary noise distributions.⁵⁶³ Such reparameterizations, however, are unknown for the Dirichlet and the logistic normal distribution. The recently introduced concrete distribution⁵⁶⁴ (simultaneously introduced as Gumbel-Softmax),⁵⁶⁵ illustrated in Fig. 7.1c, overcomes this obstacle. This distribution is supported on the simplex and parameterized by a set of deterministic variables with noise generated from stationary sources. The concrete distribution is amenable to automatic differentiation frameworks for accelerated sampling and inference. It also contains a temperature parameter, τ , which can be tuned to smoothly interpolate between the discrete categorical distribution and the uniform distribution on the simplex. As such, this temperature parameter controls the localization of constructed kernel densities towards the corners of the simplex.

Naïve GRYFFIN estimates kernel densities from concrete distributions and conditions the parameters of the concrete distribution on the sampled candidates as suggested in the PHOENICS framework. The temperature parameter is modified based on the number n of collected observations, $\tau \sim n^{-1}$, such that the priors gradually transition from a uniform distribution to a continuous approximation of the categorical distribution. Options for fu-

ture evaluations are determined *via* the acquisition function of PHOENICS, which compares the constructed kernel densities to the uniform distribution on the simplex (see Chapter 6). With a sampling parameter λ to reweight the uniform distribution, this acquisition function can favor a bias towards exploration or exploitation explicitly and natively enable batch optimization. The computational cost of the algorithm can further be reduced significantly by introducing an approximation to the computation of the kernel densities. The approximation is based on the idea that the low-density regions of the kernel densities indicate a lack of information (see Chapter 6). A precise estimate of the kernel densities in these regions might therefore not be required. We find that an approximative estimate of the kernel densities in low-density regions can significantly accelerate the computation without requiring additional evaluations. Details on this approximation are provided in the appendix of Ref. [234].

7.3.2 DESCRIPTOR-GUIDED SEARCHES WITH STATIC GRYFFIN

Naïve GRYFFIN imposes an equal measure of covariance between individual options of categorical variables, which is undesired in cases where a notion of similarity can be established between any two given options. Especially in the context of scientific discovery, similarities between the options of categorical variables can be defined, for example, *via* physicochemical descriptors for small molecules or material candidates. We extend the naïve approach by assuming that the metric to measure the similarity between any two options is based on the Euclidean distance between real-valued d -dimensional descriptor vectors, $\mathbf{x} \in \mathbb{R}^d$, which are uniquely associated with individual categorical options (see Fig. 7.1a,b).

While the descriptors are embedded in a continuous space, their arrangement in this space is unknown for a generic optimization task, and only selected points in the descriptor space can be associated with realizable categorical options. These limitations present major obstacles to optimization strategies that operate directly on the descriptor space. Instead, we propose to leverage the naïve GRYFFIN framework but redefine the metric on the simplex based on the provided descriptors. Following this strategy, the length of an infinitesimal line

element on the simplex is conditioned not only on the corresponding infinitesimal change of location on the simplex but also on the infinitesimal change of the associated descriptors. In the following, we derive an expression for the distance between any point \mathbf{z} on a simplex with dimensionality N_{opt} and a particular target corner, $\mathbf{z}_i = \delta_{it} \mathbf{e}_i$ for $i = 1, \dots, N_{\text{opt}}$, computed based on the metric spanned by the descriptors associated with individual options. The infinitesimal line element on the descriptor space with N_{desc} -many descriptors can be calculated following the Euclidean norm,

$$ds^2 = \sum_{m=1}^{N_{\text{desc}}} dx_m dx_m, \quad (7.1)$$

where we sum over all descriptors. With the assumption that descriptors are a function of the points on the simplex, $\mathbf{x} = \mathbf{x}(\mathbf{z})$, we can compute the infinitesimal changes in the descriptors via infinitesimal changes in the categorical variable,

$$dx_m = \sum_{i=1}^{N_{\text{opt}}} \frac{\partial x_m}{\partial z_i} dz_i. \quad (7.2)$$

The infinitesimal line element ds can then be expressed as

$$ds^2 = \sum_{m=1}^{N_{\text{desc}}} \sum_{i,j=1}^{N_{\text{opt}}} \frac{\partial x_m}{\partial z_i} \frac{\partial x_m}{\partial z_j} dz_i dz_j. \quad (7.3)$$

Further, if we construct $\mathbf{x}(\mathbf{z})$ as a linear function of the points on the simplex, we can replace $dx \rightarrow \Delta x$, $dz \rightarrow \Delta z$ and $ds \rightarrow \Delta s$. The length of the path from a point \mathbf{z} on the simplex to the corner \mathbf{z}_i then simplifies to

$$\Delta s^2 = \sum_{m=1}^{N_{\text{desc}}} \sum_{i,j=1}^{N_{\text{opt}}} \frac{\Delta x_m}{\Delta z_i} \frac{\Delta x_m}{\Delta z_j} \Delta z_i \Delta z_j = \sum_{m=1}^{N_{\text{desc}}} \sum_{i,j=1}^{N_{\text{opt}}} \Delta x_m^i \Delta x_m^j, \quad (7.4)$$

where we introduced Δx_m^i to describe the change of the m -th descriptor with a change in the i -th option of the categorical variable. To compute this change in the descriptor, we consider

a straight line in the categorical space,

$$\mathbf{z}(t) = (1-t)\tilde{\mathbf{z}} + t\mathbf{z}_i, \quad \text{where } t \in [0, 1], \quad (7.5)$$

and compute the value of the descriptor \mathbf{x}_m along this path. For $t = 1$, the value of \mathbf{x}_m is identical to the value of the m -th descriptor of the i -th categorical option. For $t = 0$, however, the value of \mathbf{x}_m is given by the weighted average of the descriptors x_m across all categories but the i -th category,

$$x_m(t=0) = \sum_{k \neq i}^{N_{\text{opt}}} \frac{z_k}{1-z_i} x_m^{\text{opt}_k}, \quad x_m(t=1) = x_m^{\text{opt}_i}. \quad (7.6)$$

Hence, for a given point \mathbf{z} along this path,

$$x_m(\mathbf{z}) = z_i x_m^{\text{opt}_i} + (1-z_i) \sum_{k \neq i}^{N_{\text{opt}}} \frac{z_k}{1-z_i} x_m^{\text{opt}_k}, \quad (7.7)$$

where $x_m^{\text{opt}_i}$ denotes the value of the m -th descriptor of the i -th categorical option. Following the path from \mathbf{z} to \mathbf{z}_i , we find that x_m changes as

$$\Delta x_m^i = x_m^{\text{opt}_i} - \sum_{k=1}^{N_{\text{opt}}} z_k x_m^{\text{opt}_k}. \quad (7.8)$$

We then compute the length of the path from \mathbf{z} to the corner to arrive at

$$\begin{aligned} \Delta s^2 &= \sum_{m=1}^{N_{\text{desc}}} \sum_{i,j=1}^{N_{\text{opt}}} \left(x_m^{\text{opt}_i} - \sum_{k=1}^{N_{\text{opt}}} z_k x_m^{\text{opt}_k} \right) \left(x_m^{\text{cat}_i} - \sum_{k=1}^{N_{\text{opt}}} z_k x_m^{\text{opt}_k} \right) \\ &= (N_{\text{opt}})^2 \sum_{m=1}^{N_{\text{desc}}} \left(x_m^{\text{opt}_i} - \sum_{k=1}^{N_{\text{opt}}} z_k x_m^{\text{opt}_k} \right) \left(x_m^{\text{opt}_i} - \sum_{k=1}^{N_{\text{opt}}} z_k x_m^{\text{opt}_k} \right) \end{aligned} \quad (7.9)$$

Eq. 7.9 presents the final equation to recompute distances. Based on these distances, similarities between sampled points on the simplex and its corners can be established. Kernel densities generated by the naïve approach can be transformed following this descriptor-based definition of distances on the simplex to reflect the similarity between individual options as

illustrated in Fig. 7.1a,b. As a consequence, the evaluation of one option of the categorical variable will be more informative with respect to the expected performance of other, similar options. We refer to the descriptor-guided categorical optimization as static GRYFFIN, as user-provided descriptors are used without further modifications. The benefits of static GRYFFIN over naïve GRYFFIN to accelerate the search could depend on the provided descriptors: more informative descriptors are expected to efficiently guide the algorithm to the best performing options, while less informative descriptors have the potential to mislead the algorithm. These possibilities are empirically explored in more detail in Sec. 7.4.

7.3.3 DESCRIPTOR REFINEMENT WITH DYNAMIC GRYFFIN

The dynamic formulation of GRYFFIN aims to alleviate the expected sensitivity of the performance of static GRYFFIN on the choice of provided descriptors by transforming them during the optimization. Specifically, dynamic GRYFFIN infers a transformation, T , which constructs a new set of descriptors, \mathbf{x}' , from the provided descriptors, \mathbf{x} , based on the feedback collected from evaluated options. The transformation T can be constructed to target two primary goals: (i) the generation of more informative descriptors, \mathbf{x}' , which help to navigate the candidate space more efficiently, and (ii) the interpretable identification of relevant domain knowledge to inspire design choices and scientific insights as we will demonstrate in Sec. 7.5. In addition to these two goals, the transformation T is required to be robust with respect to overfitting due to the low data scenarios which are commonly encountered in autonomous workflows.

In an attempt to balance flexibility, robustness and interpretability, we suggest to construct this transformation T from a learnable combination of the provided descriptors

$$\mathbf{x}' = \text{softsign}(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}), \quad \text{softsign}(x) = \frac{x}{1 + |x|}, \quad (7.10)$$

where \mathbf{W} and \mathbf{b} are the learnable parameters inferred from the feedback collected in previous evaluations. The class of transformations described by Eq. 7.10 includes slightly non-linear

translations and rotations of the provided descriptors. While more complex transformations accounting for higher-order interactions between individual descriptors could potentially yield even more informative descriptors, slightly non-linear transformations are inherently robust to overfitting⁵⁶⁶ and more amenable to intuitive interpretation than more complex models.⁵⁶⁷ We will demonstrate empirically in Secs. 7.4 and 7.5 that this class of transformations is well suited for a variety of categorical optimization tasks. Following a stochastic gradient optimization, the parameters \mathbf{W} and \mathbf{b} in Eq. 7.10 are adjusted to (i) increase the correlation between the newly generated descriptors \mathbf{x}' and the associated measurements, (ii) reduce correlations between newly generated descriptors, and (iii) remove redundant descriptors with poor correlations with the measurements or high correlations with other newly generated descriptors. These three goals are modeled as penalties which are to be minimized at training time (see appendix of Ref. [234] for details).

7.4 SYNTHETIC BENCHMARKS

We empirically assess the performance of the introduced variants of GRYFFIN on a set of synthetic benchmark surfaces (see appendix of Ref. [234] for details). Four of the surfaces constitute categorized adaptations of established functions commonly used to benchmark global and local optimization strategies on continuous parameter domains. We also include three partially and fully randomized surfaces with responses sampled from stationary probability distributions. While the ordering of the categorical options is arbitrary, we introduce a reference order to illustrate the surfaces (see appendix of Ref. [234] for details). Unless noted otherwise, descriptors for the categorical options are generated such that they encode the reference order. Implementations of all benchmark surfaces are made available on GitHub.⁷⁹ GRYFFIN is compared to a set of qualitatively different optimization strategies which are implemented in publicly available libraries: genetic optimization through `PYEVOLVE`,^{199,200,568} Bayesian optimization with random forests (RFs) as implemented in `SMAC`,^{241–243} Bayesian optimization with Gaussian processes (GPs) *via* `GPYOPT`,^{516,558,569,570} and Bayesian optimization with tree-structured Parzen windows introduced in `HYPEROPT`.^{505,571} We also run

random explorations of the candidate space as a baseline. We compare the performance of the different formulations of GRYFFIN to the other optimization strategies on all benchmark surfaces, probe the influence of the number of descriptors, study the scaling of GRYFFIN with the number of options per categorical variable and the number of categorical variables, and investigate the benefits of dynamic GRYFFIN over static GRYFFIN. For all comparisons, we measure the fraction of the candidate space that a given optimization strategy explored to locate the best candidate. Unless noted otherwise, all comparisons are averaged over 200 independent executions.

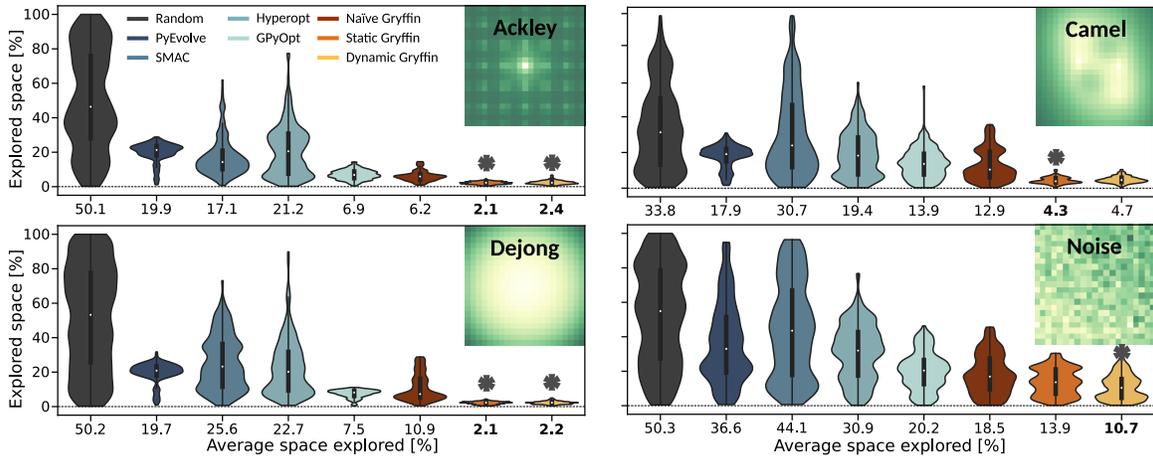


Figure 7.2: Performances of various optimization strategies on selected synthetic surfaces. Individual panels indicate the fraction of the candidate space each algorithm explored before finding the global minimum of the surfaces illustrated on the top right (low values are shown in yellow, high values in green), averaged over 200 independent executions. Best performing algorithms are indicated in bold by a star. Color codes for the optimization strategies are shown in the top panel and apply to all panels in this figure. Reproduced with permission from Ref. [234].

7.4.1 OPTIMIZATION PERFORMANCE

In a first test, we compare the optimization strategies on two-dimensional formulations of the synthetic benchmark surfaces with 21 options per dimension, as illustrated in Fig. 7.2. The *Dejong* surface generalizes the convex parabola from continuous to categorical spaces, such that we consider it to be *pseudo-convex*. In contrast, the *Ackley* surface is generalized from the Ackley path function, which is non-convex on the parameter domain. The *Camel* surface presents a degenerate global optimum as there are two different combinations of

options that yield the same optimal response. Finally, the reference ordering of the *Noise* surface arranges the options such that they yield a relatively large local relative variance.[†] The fractions of the surfaces that the optimizers explored to locate their optima, averaged over 200 independent executions, are illustrated in Fig. 7.2. Full optimization traces are reported in the appendix of Ref. [234]. We observe that a random exploration of the space requires the evaluation of approximately half the space for the surfaces with well-defined global optima, and about a third of the space for the *Camel* function with a singly degenerate optimum. We find that the performances of PYEVOLVE, SMAC, and HYPEROPT are roughly comparable across the different surfaces, although PYEVOLVE tends to outperform SMAC and HYPEROPT on the noiseless surfaces. GPYOPT generally locates global optima faster than the other strategies but is slightly slower than naïve GRYFFIN on the non-convex surfaces. The faster optimization of convex surfaces with GP-based Bayesian optimization compared to kernel density augmented Bayesian optimization has already been observed and discussed for continuous domains (see Chapter 6).²³³ Notably, the static and dynamic formulations of GRYFFIN can significantly outperform the other optimization strategies, with reductions of the explored space by several factors. This observation confirms that providing real-valued descriptors can substantially accelerate the search. We also observe similar performances of the static and dynamic formulations of GRYFFIN for the deterministic surfaces (*Ackley*, *Dejong*, *Camel*), while dynamic GRYFFIN optimizes the noisy surface at a faster rate. This observation suggests that dynamic GRYFFIN is indeed capable of learning more informative descriptors.

7.4.2 SCALING TO MORE OPTIONS AND HIGHER DIMENSIONS

We further study the performance of GRYFFIN for larger candidate spaces with (i) more categorical variables, and (ii) more options per categorical variable. Increasing the number of variables or the number of options per variable generally increases the number of candidates in the space and is thus expected to require more evaluations overall before the best can-

[†] Note, that the locality of the variance in the response is measured with respect to the descriptor vectors of the associated options.

didate is identified. Results obtained for these benchmarks are detailed in the appendix of Ref. [234]. The benchmarks suggest that GRYFFIN indeed uses more candidate evaluations to locate the global optimum with an increasing volume of the search space consistently across all benchmark surfaces. However, although the number of evaluations increases, the fraction of the explored space generally decreases. More specifically, we identify a polynomial decay of the explored space with an increasing number of options per variable, with decay exponents ranging from -1.0 to -1.25 and an exponential decay for an increase in the number of parameters with decay coefficients ranging from -1.6 to -2.0 across the different surfaces as shown in more detail in the appendix of Ref. [234]. Based on this observation, we conclude that GRYFFIN may show an onset of the *curse of dimensionality*⁵⁷² only for a relatively large number of dimensions and indeed constitutes an optimization strategy which can navigate large categorical spaces efficiently.

7.4.3 DATA-DRIVEN REFINEMENT OF DESCRIPTORS

The effectiveness of transforming provided descriptors to accelerate the search is studied in detail on the *slope* surface with 51 options per dimension, resulting in 2,601 different candidates (see Fig. 7.3). For this benchmark, we randomly assign descriptors to each of the categorical options at a desired targeted correlation between the descriptors and the responses of the associated options. With a decreasing correlation, the local variance increases, which results in a less structured space that is more challenging to navigate. We therefore generally expect a performance degradation for both static and dynamic GRYFFIN with decreasing correlation. Fig. 7.3 illustrates the fractions of the candidate space explored by static and dynamic GRYFFIN for different targeted correlations between the supplied descriptors and the responses. For comparison, we also report the performance of the descriptor-less naïve formulation of GRYFFIN, which is independent of the supplied descriptors. We observe a significant increase in the fraction of the explored space with a decreasing correlation for both static and dynamic GRYFFIN. Although both methods require more candidate evaluations with less informative descriptors, their performance never significantly degrades beyond the

naïve formulation, indicating that even entirely uninformative descriptors do not delay the search for the best candidate compared to descriptor-less scenarios.

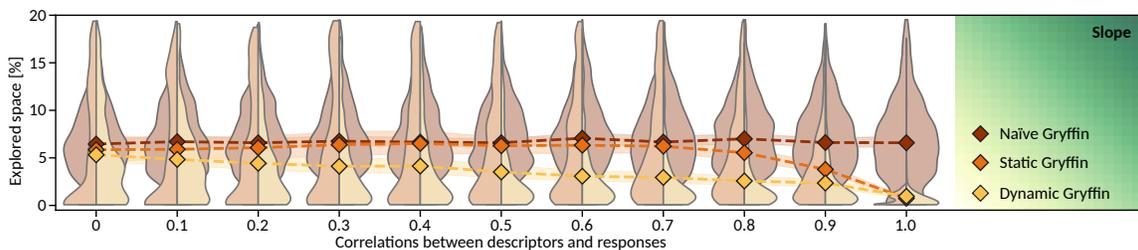


Figure 7.3: Behavior of naïve, static and dynamic GRYFFIN on the slope surface for different descriptors. Descriptors have been generated randomly with a targeted correlation between the descriptors and the responses of the associated options. Reproduced with permission from Ref. [234].

We further find that static GRYFFIN can benefit from descriptors and significantly outperform the naïve approach if the Pearson correlation coefficient between descriptors and responses is at least 0.8. Below this value, the average performance of static and naïve GRYFFIN is comparable, although the variance on the performance is higher for the static formulation. Similar to static GRYFFIN, learning a more informative set of descriptors with dynamic GRYFFIN accelerates the search more if the correlation between the descriptors and the responses is high. However, the dynamic formulation is generally at least as fast as the static formulation and can successfully leverage descriptors to outperform descriptor-less searches even at correlations as low as 0.1. We thus confirm that the descriptor transformation introduced in Eq. 7.10 is sufficiently robust to be applied to low-data tasks and conclude that deploying dynamic GRYFFIN can be beneficial for some descriptor guided optimization tasks without delaying the optimization compared to static GRYFFIN.

7.5 APPLICABILITY OF GRYFFIN TO CHEMISTRY AND MATERIALS SCIENCE

Following the empirical benchmarks of GRYFFIN, we now demonstrate its applicability and practical relevance to a set of optimization tasks across materials science and chemistry. Specifically, we target the discovery of non-fullerene acceptors for OSCs, the design of HOIPs for light-harvesting, and the selection of phosphine ligands simultaneously to the optimization of process conditions for Suzuki-Miyaura coupling reactions. Obtaining statistically

significant performance comparisons at a sufficient level of confidence requires the repeated execution of optimization runs to average out the influences of initial conditions and probabilistic elements of the optimization strategies. As repetitive executions of optimization runs on these applications would be highly resource-demanding experimentally or computationally, we construct these optimization tasks from recently reported datasets. The discovery of non-fullerene acceptors and perovskites is based on lookup tables, and the optimization of Suzuki reactions is facilitated *via* a probabilistic model trained on experimental data (*virtual robot*) to emulate experimental uncertainties in addition to the average response. The concept of virtual robots to benchmark experiment planning strategies is discussed in more detail in Chapter 8. The selection of physicochemical descriptors for the categorical variables of these three applications is mostly motivated by their accessibility. Determining the most suitable set of descriptors for a given application requires repeated measurements of the property of interest, which typically is a highly resource-demanding process. As such, the selection of the descriptors as outlined in the following sections balances applicability and availability, as the most informative set of descriptors for a scientific discovery task might be *a priori* unknown.

7.5.1 DISCOVERY OF NON-FULLERENE ACCEPTOR CANDIDATES FOR ORGANIC PHOTOVOLTAICS

Small organic molecules currently constitute the highest performing acceptor materials for OSCs (see Sec. 2.1).^{156,157} The large number of degrees of freedom when designing such non-fullerene acceptors, arising from complex aromatic molecular geometries, allows us to fine-tune their relevant electronic properties, *e.g.*, the optical gap and the energy level alignment between the donor and acceptor materials. Despite their flexibility, the large design space for non-fullerene acceptors poses significant obstacles to the discovery of promising candidate molecules.

We demonstrate the applicability of GRYFFIN for the discovery of non-fullerene acceptors on a candidate space of 4,216 different small organic molecules, which form a subset of a re-

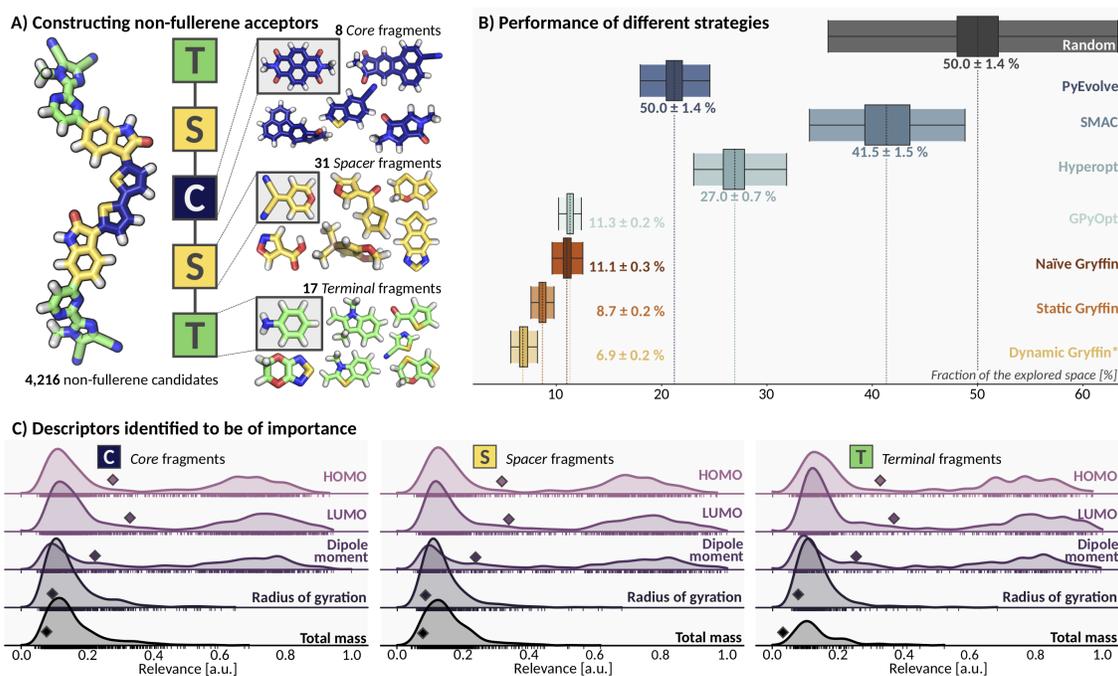


Figure 7.4: Performance of GRYFFIN on the task of identifying non-fullerene acceptors for maximized power conversion efficiencies. (A) Non-fullerene acceptor candidates are constructed from a set of molecular fragments (core, spacer and terminal) which are symmetrically arranged to span a library of 4,216 different candidate molecules. (B) Fraction of the candidate library to be explored by each of the studied optimization strategies to identify the best performing acceptor candidate. (C) Most informative descriptors to guide the search of *dynamic* GRYFFIN. Diamonds indicate the average relevance of each descriptor. Reproduced with permission from Ref. [234].

cently reported comprehensive study.⁴⁹⁷ Acceptor candidates in this library are constructed from a set of molecular fragments that are separated into three fragment pools (see Fig. 7.4a). Each candidate is composed of one core fragment *C* (8 options), two spacer fragments *S* (31 options), and two terminal fragments *T* (17 options) following a symmetric design. Details on the library of candidate fragments are reported in the appendix of Ref. [234]. The performance of each acceptor candidate is quantified based on the power conversion efficiency (PCE) which is computed *via* the Scharber model following a workflow based on the data-driven calibration of density functional theory (DFT) results.⁴⁹⁷ The optimization task targets the maximization of the expected PCE of the acceptor candidate. We guide static and dynamic GRYFFIN with a set of electronic and geometric descriptors for each of the fragments: the HOMO and LUMO energy levels, the dipole moment, the radius of gyration and the molecular weight. Electronic properties were computed at the B3LYP/Def2SVP level

of theory on a SuperFineGrid using Gaussian,⁴²¹ and the radius of gyration was computed for the ground state conformations of the molecules. The correlations of the descriptors with the PCE of the resulting non-fullerene acceptor are generally low, with the highest encountered Pearson correlation coefficients reaching values of about 0.2 (see appendix of Ref. [234] for details). In fact, the identification of improved descriptors for the accurate prediction of PCE in OSCs is an active field of research.⁵⁷³⁻⁵⁷⁵

Fig. 7.4b illustrates the fraction of the candidate library averaged over 200 independent executions, which each optimization strategy explored before identifying the combination of fragments which yields the highest PCE. Full optimization traces for each of the optimization strategies are reported in the appendix of Ref. [234]. In agreement with the synthetic tests (see Sec. 7.4), we find that PYEVOLVE explores smaller fractions (21 %) of the space than HYPEROPT (27 %) or SMAC (41 %). The performance of naïve GRYFFIN, exploring about 11 %, is comparable to GPYOPT and thus significantly faster than the other benchmark strategies. However, the physical descriptors supplied for each of the fragments enable static GRYFFIN to find the best acceptor candidate after exploring only 8.7 % of the candidate space (~ 22 % reduction of the required acceptor evaluations). In contrast, dynamic GRYFFIN can refine the supplied descriptors to find the best candidate with only 6.9 % of the library explored (~ 38 % reduction over naïve search). This improvement of dynamic GRYFFIN over static GRYFFIN confirms that the supplied descriptors can be transformed into a more informative set to accelerate the search. Fig. 7.4c illustrates the importance of individual descriptors to guide the search, as determined by *dynamic* GRYFFIN. Specifically, we plot the relative contributions of individual descriptors to the set of transformed descriptors that were used when the best performing candidate was identified. We observe that the descriptor search emphasizes the relevance of electronic descriptors over geometric descriptors consistently across all types of fragments. Indeed, the Scharber model is designed to estimate power conversion efficiencies qualitatively from the electronic properties of the acceptor material,^{575,576} and further refinements can be enabled by considering the bandgap. The design of non-fullerene acceptor candidates beyond the provided library could, therefore,

be inspired mostly by the electronic properties of the fragments rather than their geometric properties, although even more informative descriptors could potentially be constructed with more computational effort.⁵⁷³

7.5.2 DISCOVERY OF HYBRID ORGANIC-INORGANIC PEROVSKITES FOR LIGHT-HARVESTING

perovskite solar cells (PSCs) constitute another class of third-generation light-harvesting materials which are typically composed of inorganic lead halide matrices and contain inorganic or organic anions (see Fig. 7.5a).^{151–153} Recently, PSCs have experienced increased attention as breakthroughs in materials, and device architectures boosted their efficiencies and stabilities.¹⁵⁴ Nevertheless, the discovery of viable perovskite designs involves numerous choices regarding material compositions and process parameters, which poses a challenge to the rapid advancement of this light-harvesting technology. This second demonstration of the applicability of GRYFFIN focuses on the discovery of HOIPs based on a recently reported dataset.⁵⁷⁷ The HOIP candidates of this dataset are designed from a set of four different halide anions, three different group-IV cations, and 16 different organic anions, resulting in 192 different HOIP compositions. Among other properties, the dataset reports the bandgaps of the HOIP candidates obtained from DFT calculations with GGA and the HSE06 functional. In this application, we aim to minimize the bandgap.

The inorganic constituents are characterized by their electron affinity, ionization energy, mass, and electronegativity to guide the searches of the static and dynamic formulations of GRYFFIN. The organic compounds are described by their HOMO and LUMO energy levels, dipole moment, atomization energy, radius of gyration, and molecular weight. All electronic descriptors were computed at the HSEH1PBE/Def2QZVPP level of theory on a SuperFineGrid with Gaussian,⁴²¹ and the radii of gyration are calculated for the ground state conformer. Note, that in contrast to the search for viable non-fullerene acceptors, this application presents an optimization task which not only features physically different descriptors between individual categorical variables but also varying dimensionalities of the descriptors associated with individual categorical variables. However, the correlations be-

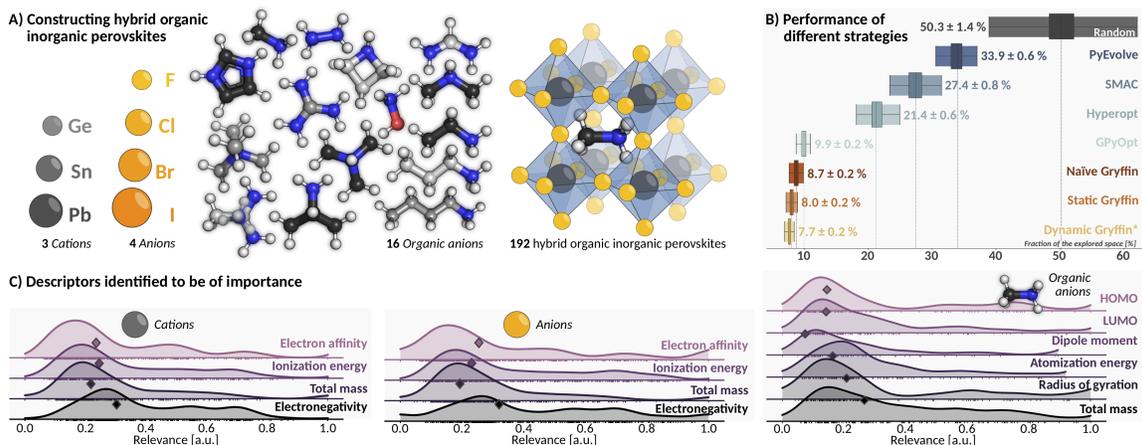


Figure 7.5: Results of the benchmarks on hybrid organic inorganic perovskites. (A) Perovskites are assembled by choosing one of three inorganic cations, one of four inorganic anions and one of 16 organic anions, resulting in 192 unique designs. (B) Fractions of the candidate library to be explored by each of the studied optimization strategies to identify the perovskite design with the lowest bandgap. (C) Most informative descriptors to guide the optimization for each of the three constituents identified by dynamic GRYFFIN. Diamonds indicate the average relevance of each descriptor. Reproduced with permission from Ref. [234].

tween individual descriptors and the expected bandgaps of the assembled HOIP materials are significantly higher compared to the descriptors used for the non-fullerene acceptors (see appendix of Ref. [234] for additional details).

The fractions of the candidate space explored by each optimization strategy before locating the HOIP composition with the lowest bandgap are illustrated in Fig. 7.5b. More detailed results are reported in the appendix of Ref. [234]. Similarly to the synthetic benchmarks (see Sec. 7.4) and the optimization of non-fullerene acceptors (see Sec. 7.5.1), we find that all optimization strategies outperform a purely random exploration of the candidate space. Bayesian optimization approaches tend to locate the best performing HOIP candidate at a faster rate than PYEVOLVE ($\sim 34\%$), with GPYOPT evaluating only about 10% of the candidate space, followed by HYPEROPT ($\sim 21\%$) and SMAC ($\sim 27\%$). However, naïve GRYFFIN succeeds after exploring less than 9% of the search space. Note that this fraction of the search space corresponds to roughly 17 HOIP candidates, which approximately matches the number of available organic compounds. The static and dynamic formulations of GRYFFIN even undercut this value and identify the best performing HOIP within less than 8%, corresponding to the evaluation of less than 16 HOIP candidates on average.

This observation confirms that GRYFFIN indeed accelerates the optimization of categorical variables if physical descriptors are available. However, we no longer observe a significant performance difference between the static and the dynamic formulation of GRYFFIN, which is in agreement with the significantly higher correlation between provided descriptors and the bandgaps (see Sec. 7.4.3). For this application, we find that electronegativity is most relevant for the inorganic constituents, while the radius of gyration and the molecular weight are most informative for the organic compound. Although the targeted property (bandgap of the HOIP) is an electronic property, dynamic GRYFFIN seems to benefit the most from the geometric and not the electronic descriptors of the organic compound. In contrast, the mass of the inorganic compounds seems to be the least relevant, while their electronegativity is most informative. These observations suggest that the organic molecule does not directly affect the electronic properties of the HOIP material but rather induces a change in the arrangement of the inorganic compounds, which in turn modulates the bandgap. Indeed, this hypothesis has emerged in various studies on perovskite materials,^{578–582} which confirms that dynamic GRYFFIN can capture the relevant trends in the descriptors and inspire future design choices.

7.5.3 SUZUKI-MIYAUURA CROSS-COUPPLING OPTIMIZATION

As a final application, we demonstrate how GRYFFIN can aid in the optimization of Suzuki-Miyaura cross-coupling reactions with heterocyclic substrates.⁵⁸³ These reactions are of particular interest to the pharmaceutical industry,⁵⁸⁴ and have recently been studied in the context of self-optimizing reactors for flow chemistry.^{191,553,554} The optimization of chemical reactions typically targets a maximization of the yield. The yield of a reaction can be modified by varying a set of process conditions, which can largely be described by continuous variables. However, the yield can also be increased by using suitable catalytic systems that modulate the rate of the reaction. For the example of a flow-based Suzuki-Miyaura cross-coupling reaction, we specifically consider three continuous reaction conditions (temperature, residence time, catalyst loading) and one categorical variable (ligand for palladium catalyst)

as illustrated in Fig. 7.6a. The optimization task targets the maximization of the reaction yield, while simultaneously maximizing the turnover number (TON) of the catalyst. We employ the CHIMERA scalarizing strategy²³⁵ to enable this multi-objective optimization (see Chapter 8), where we accept a 10% degradation on the maximum achievable reaction yield to increase the TON as the secondary objective. This acceptable degradation corresponds to the desired reaction yield of above 85.4%. GRYFFIN is integrated with the PHOENICS algorithm to optimize categorical and continuous parameters simultaneously (see Chapter 6). We consider a set of seven ligands (see Fig. 7.6b), which are characterized by their molecular weight, the number of rotatable bonds, their melting points, the number of valence electrons and their partition coefficients, quantified by $\log P$. Details on the physicochemical descriptors and the ranges for the continuous parameters are provided in the appendix of Ref. [234].

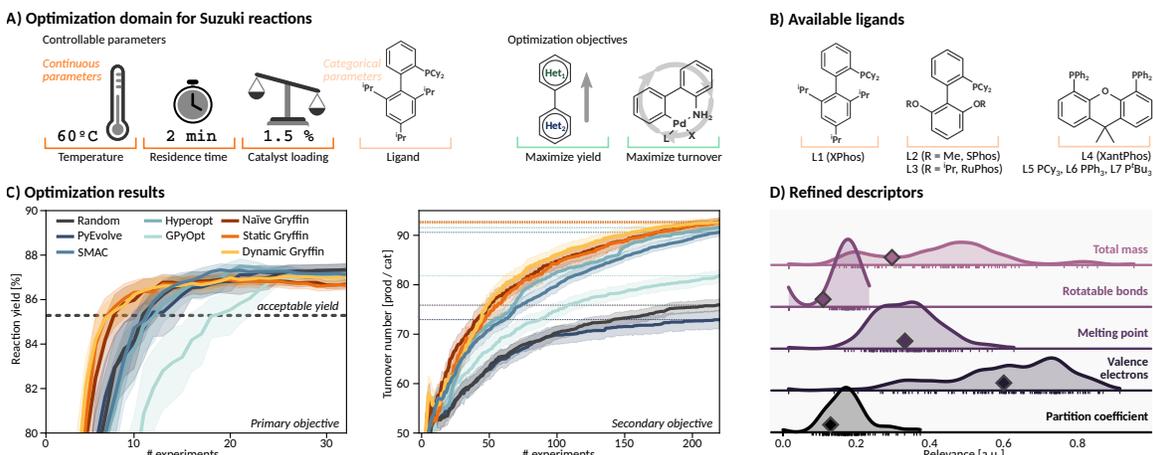


Figure 7.6: Results of the benchmarks on the optimization of Suzuki-Miyaura reactions. (A) Parameters and objectives for the optimization task: We target the identification of optimal values for three process parameters (temperature, residence time, catalyst loading) and one categorical variable (ligand) to maximize the yield of the reaction and the turnover number (TON) of the catalyst. (B) Illustration of the seven available ligands. (C) Optimization traces showing the performance of individual optimization strategies. (D) Most informative descriptors to guide the optimization. Diamonds indicate the average relevance of each descriptor. Reproduced with permission from Ref. [234].

As complete performance analyses of the different optimization strategies are experimentally not tractable, we emulate noisy experimental responses with probabilistic models which are trained on experimental data, as previously demonstrated for benchmarking multi-objective optimization strategies.²³⁵ Specifically, we train a Bayesian neural network (BNN)

to reproduce the reaction yield and TON of a previously reported flow-based reactor.¹⁹¹ Details on the data acquisition, model training, and prediction accuracies are provided in the appendix of Ref. [234]. The coefficients of determination of the trained model for the test set predictions of the reaction yield and the TON are above $r^2 > 0.96$, which indicates that the trained model indeed constitutes a realistic approximation to the experimental surface. With this experimental emulator, we execute 200 independent optimization runs with 240 evaluations for each of the benchmarked experiment planning strategies.

Fig. 7.6c illustrates the performance of the individual optimization strategies. We find that GPYOPT requires about 18 evaluations to identify reaction conditions which achieve desired yields of above 85.4%. The other benchmark strategies, including random exploration, satisfy this first objective already after evaluating approximately 10-12 different conditions. Only the three formulations of GRYFFIN can locate desired reaction conditions at an even faster rate, requiring 7-8 evaluations. For the subsequent maximization of the TON, we observe that PYEVOLVE is the slowest of the optimization strategies. In fact, random search outperforms PYEVOLVE after about 100 evaluations. Despite its relatively poor performance for the reaction yield, GPYOPT maximizes the TON faster than random search. However, SMAC and HYPEROPT still achieve significantly higher TONs for any given number of evaluations and are slightly outperformed by GRYFFIN. We do not observe a significant difference in the performance of the three formulations of GRYFFIN, which can be attributed to the fact that we only have one categorical variable with only seven options. Nevertheless, we observe a slight trend that dynamic GRYFFIN achieves slightly higher TONs than static or naïve GRYFFIN.

The contributions of individual descriptors are illustrated in Fig. 7.6d, where we find that the number of valence electrons shows the highest relevance among all descriptors to guide dynamic GRYFFIN, while the number of rotatable bonds is the least relevant. Indeed, the number of rotatable bonds correlates the least with the maximum and average reaction yields and TONs for any values of the other parameters (see appendix of Ref. [234]), confirming that dynamic GRYFFIN correctly identifies non-informative descriptors within

the given library of ligand candidates. The number of valence electrons also correlates strongly with the maximum achievable reaction yield for each of the ligands, confirming that this descriptor is highly informative to identify ligands which satisfy the reaction yield threshold. Melting point and molecular weight are likely indicated as relevant due to their strong correlation with the number of valence electrons. Based on the indications of dynamic GRYFFIN, the design of more potent ligand candidates could be inspired by the number of valence electrons. However, it is essential to mention that here, in contrast to the other applications, we considered a relatively small library of only seven ligands, such that the descriptor indications might not necessarily generalize well to more extensive libraries.

Overall, across all three applications, we find that naïve GRYFFIN constitutes a competitive strategy for the optimization of categorical variables in chemistry and materials science, which tends to outperform state-of-the-art optimization strategies without leveraging physicochemical descriptors in the selection process. Static GRYFFIN can accelerate the search with provided descriptors and navigate the search space more efficiently by exploiting descriptor-based similarities between individual options, thus leveraging domain knowledge. Dynamic GRYFFIN can accelerate the search even further by transforming provided descriptors to improve their relevance and inspire scientific insights. Finally, GRYFFIN integrates well with optimization strategies for continuous variables and enables the simultaneous optimization of mixed continuous-categorical parameter spaces.

7.6 CONCLUSION

In this chapter, we introduced GRYFFIN, a data-driven experiment planning strategy for the selection of categorical variables such as functional molecules, catalysts, or material constituents in autonomous experimentation workflows. GRYFFIN can leverage domain knowledge in the form of physicochemical descriptors for each of the categorical options, and inspire design choices and scientific insights while efficiently navigating the search space. To this end, GRYFFIN is based on the idea to augment Bayesian optimization with kernel density estimation, which has recently been introduced for continuous optimization domains

(see Chapter 6).²³³ Using smooth approximations to categorical distributions and locally transforming the metric of the optimization domain, GRYFFIN can exploit similarity information between categorical options to accelerate the search for promising molecules and materials.

We assessed the performance of GRYFFIN in comparison to state-of-the-art strategies to select categorical variables on a set of synthetic benchmark functions. Our benchmarks indicate that the naïve formulation of GRYFFIN, which does not use any descriptor information, is competitive on *pseudo-convex* surfaces and outperforms the other strategies on all other surfaces. Descriptor-guided searches with static GRYFFIN identify global optima at significantly faster rates consistently for all surfaces. Dynamic GRYFFIN, which attempts to construct a more informative set of descriptors, can accelerate the search even further in some cases, especially for moderate correlations between the descriptors and the responses and for noisy environments. The capabilities of GRYFFIN were further demonstrated on three real-world applications across materials science and chemistry: (i) the discovery of non-fullerene acceptors for OSCs, (ii) the discovery of HOIPs for light-harvesting, and (iii) the mixed categorical-continuous selection of ligands and reaction conditions for Suzuki-Miyaura cross-coupling reactions. GRYFFIN outperforms the other experiment planning strategies in all three applications. Static and dynamic GRYFFIN can accelerate the searches even with moderately informative physicochemical descriptors. We further find that dynamic GRYFFIN can identify trends among the descriptors which elucidate some of the prevalent phenomena which give rise to the properties of interest, indicating that dynamic GRYFFIN has the potential to foster scientific understanding and encourage physical and chemical intuition for the studied systems.

Based on the synthetic and real-world benchmarks, we suggest that GRYFFIN constitutes a readily available strategy for the efficient selection of categorical variables in data-driven experimentation workflows and alleviates some of the immediate challenges to the versatile deployment of autonomous experimentation platforms. The demonstrated acceleration of the search based on physicochemical descriptors constitutes a step towards autonomous

experimentation guided by domain knowledge. In summary, we believe that GRYFFIN has the potential to accelerate scientific discovery and invite the community to test and deploy it to scenarios where evaluations of categorical parameters are expensive, and similarities between categorical options can be defined.

8

Chimera: enabling hierarchy-based multi-objective optimization for self-driving laboratories

Apart from minor modifications, this chapter was originally published by the Royal Society of Chemistry as:

Chimera: enabling hierarchy-based multi-objective optimization for self-driving laboratories. Florian Häse, Loïc M. Roch and Alán Aspuru-Guzik, *Chem. Sci.* **9** (39), 7642–7655.

Reproduced from Ref. [235] with permission from the Royal Society of Chemistry.

ABSTRACT

Finding the ideal conditions satisfying multiple desired targets simultaneously is a challenging decision-making process, which impacts science, engineering, and economics. Additional complexity arises for tasks involving experimentation or expensive computations, as the number of affordable evaluations is restricted by a budget. We propose CHIMERA as a general-purpose achievement scalarizing function for multi-target optimization where evaluations are the limiting factor. CHIMERA combines concepts of *a priori* scalarizing with lexicographic approaches and is applicable to any set of n unknown objectives. Importantly, it does not require detailed prior knowledge about individual objectives. The performance of CHIMERA is demonstrated on several well-established synthetic multi-objective benchmark sets using different single-objective optimization algorithms. We further illustrate the applicability and performance of CHIMERA on two practical examples: (i) the auto-calibration of a virtual robotic sampling sequence for direct-injection, and (ii) the inverse-design of a four-pigment excitonic system for efficient energy transport. The results suggest that CHIMERA

enables a broad class of optimization algorithms to rapidly find ideal conditions. Additionally, the presented applications highlight the interpretability of CHIMERA to corroborate design choices on tailoring system parameters.

8.1 MULTI-OBJECTIVE OPTIMIZATION FOR SCIENTIFIC DISCOVERY

Multi-objective optimization is ubiquitous across various fields in science, engineering, and economics. It can be interpreted as a multi-target decision-making process,⁵⁸⁵ aiming at finding the ideal set of conditions, *e.g.*, parameters of experimental procedures, theoretical models, or computational frameworks, which yield the desired pre-defined targets. In chemistry and material science, these targets can include the yield and selectivity of reactions, the power conversion efficiency (PCE) and photostability of a solar cell, production cost and overall execution time of processes, or optimization of materials with properties tailored to specific needs. In general, the ideal conditions for which all targets assume their desired optimal values do not exist. Improving on one target might only be possible at the expense of degrading on other targets (see Sec. 2.3.2). Straightforward approaches to determine ideal conditions satisfying multiple targets are formulated as detailed systematic searches of all possible conditions. However, these strategies require numerous objective evaluations, scale exponentially with the number of conditions to be optimized, and do not guarantee to locate the ideal conditions. Applications involving experimentation or expensive computations are, therefore, beyond the viability of these searches as the number of conducted experiments or computations must be kept low. Instead, robust and efficient algorithms evolving on multi-dimensional surfaces are needed to identify optimal conditions within a minimum number of distinct evaluations.

Such robust and efficient algorithms have the potential to open new avenues to multi-objective optimization in chemistry and materials science when combined with autonomous experimentation as implemented in self-driving laboratories (see Part III). Such laboratories combine artificial intelligence (AI) with automation, and enable the design and execution of experiments in full autonomy, without human interaction.^{203,417,508,586–588} The learning

procedure suggests new conditions while accounting for the observed merit of previously conducted experiments, forming a closed-loop. Consequently, self-driving laboratories learn experimental conditions on-the-fly by continuously refining parameters to maximize the merit of the machine-proposed conditions and satisfy pre-defined targets.^{321,589} However, applications with multiple objectives pose the challenge of formulating an optimal solution based on tolerated trade-offs in the objectives. To address this challenge, competing criteria need to be balanced to identify the conditions which yield the highest merit under user-defined preferences. Hereafter, we propose CHIMERA, a versatile achievement-scalarizing function (ASF) for multi-objective optimization with costly to evaluate objectives.

Recently, multi-objective optimization strategies have been successfully applied to various scenarios. Examples include the rational design of dielectric nanoantennas,⁵⁹⁰ and plasmonic waveguides,⁵⁹¹ the optimization of Stirling heat pumps,⁵⁹² the design of thermal energy storage systems,⁵⁹³⁻⁵⁹⁵ and optimizations on scheduling problems in combined hydro-thermo-wind power plants.⁵⁹⁶ However, in the applications mentioned above, the merit of a set of conditions could be assessed by analytic models, which were fast to evaluate computationally. As such, these optimization tasks could be approached with methods identifying the entire set of solutions that cannot be further optimized in at least one of the objectives at the expense of numerous objective evaluations. Preference information regarding specific solutions could then be expressed, knowing the surface of optimal points. In chemistry, multi-objective optimization methods have been applied to determine trade-offs in the reaction rate and yield of methylated ethers,⁵⁹⁷ maximize the intensity of quantum dots at a target wavelength,⁵⁹⁸ or balance the production rate and conversion efficiency of Paal-Knorr reactions.⁵⁹⁹ These optimization tasks have been approached with methods that allow expressing preference information before starting the optimization procedures. Preference information was provided by constructing a single merit-function from all considered objectives such that the single merit-based function accounts for the provided preferences. Optimizations were then conducted on the merit-based function using single-objective optimization algorithms.

The abovementioned examples display the successful application and benefit of multi-

objective optimization methods on self-optimizing reactors, illustrating how they can power self-driving laboratories. Yet, the merit-based functions employed in these examples are often handcrafted. Constructing a suitable and versatile merit-based function with little prior knowledge about the objectives is challenging.^{503,600} As a matter of fact, compositions of merit-based functions can sometimes require refinements after initial optimization runs as the desired preference in the objectives is not achieved.⁵⁹⁹ Recently, Walker *et al.* have introduced a framework for formulating merit-based multi-objective optimization as constrained optimization problems for the synthesis of o-xylenyl adducts of Buckminsterfullerenes.¹⁹² Their approach aims to optimize a main objective while keeping other objectives at desired levels by considering them as constraints. However, their method depends on the choice of constraints, which requires substantial prior knowledge about the objective surfaces. The lack of a universal, general-purpose method for constructing merit-based functions from multiple objectives is a challenge to design problems and appears as a significant obstacle to the massive deployment of self-optimizing reactors and self-driving laboratories. Notably, we identify two main constraints based on the considerations in Sec. 1.2: (i) objective evaluations involve timely and costly evaluations (experimentally or computationally), and, thus, must be kept to a minimum, (ii) no prior knowledge is available about the surface of the objectives. In this work, we use these constraints as requirements for the formulation of CHIMERA.

CHIMERA is an approach to multi-objective optimization for experimental and computational design. It combines concepts of *a priori* scalarizing with ideas from lexicographic approaches and is made available on GitHub.⁸⁰ Hereafter, we show on several well-established benchmark sets and on two practical applications how CHIMERA fulfills the constraints mentioned before. Our proposed method relies on preference information provided in the form of a hierarchy in the objectives. A single merit-based function is constructed from the provided hierarchy and shapes a surface which can be optimized by a variety of single-objective optimization algorithms. CHIMERA does not require detailed assumptions about the surfaces of the objective functions, and it improves on the hierarchy of objectives from the beginning

of the optimization procedure, without any required warm-up iterations.

8.2 FORMULATING CHIMERA

We consider a Pareto optimization problem with n objective functions $\{f_k\}_{k=0}^{n-1}$ defined on the d -dimensional compact subset $\mathcal{P} \subset \mathbb{R}^d$. Sec. 2.3.2 provides a comprehensive discussion on the background of multi-objective optimization. We further assume that no prior information about the objectives is available and that evaluations of the objectives are demanding in terms of budgeted resources, motivating *a priori* methods with gradient-free global optimization algorithms. In this section, we detail CHIMERA, which follows the idea of lexicographic approaches by providing preference information in the form of a hierarchy in the objectives, but formulates a single ASF based on the provided hierarchy (see Fig. 8.1). The formulation of CHIMERA enables the following procedure: (i) Given a hierarchy in the objectives, relative tolerances are defined for each objective, indicating the allowed relative deviation with respect to the full range of the objective values. (ii) Improvements to the main objective should always be realized unless sub-objectives can be improved without degrading the main objective beyond the defined tolerance. (iii) Changes in the order of the hierarchy and the tolerances on the objectives should enable the optimization procedure to reach different Pareto optimal points. Cases, where two or more objectives are judged to be of equal importance, can be accounted for by combining these objectives into a single objective.

8.2.1 CONSTRUCTING CHIMERA

We assume the set of $\mathbf{f} = (f_0, \dots, f_{n-1})$ objective functions to be ordered based on a descending hierarchy, *i.e.*, f_0 is the main objective, and that the optimization procedure aims to minimize each of the objectives. An example of a set of three objective functions is illustrated in Fig. 8.1a. CHIMERA is updated at every optimization iteration based on all available observed pairs of parameter points and objectives $\mathcal{D}_j = \{(\mathbf{x}_i, \mathbf{f}_i)\}_{i=1}^j$. Using prior observations \mathcal{D}_j , relative tolerances \tilde{f}_k^{tol} defined prior to the optimization procedure are used to compute absolute tolerances f_k^{tol} on all objectives at each optimization iteration (see Eq. 8.1).

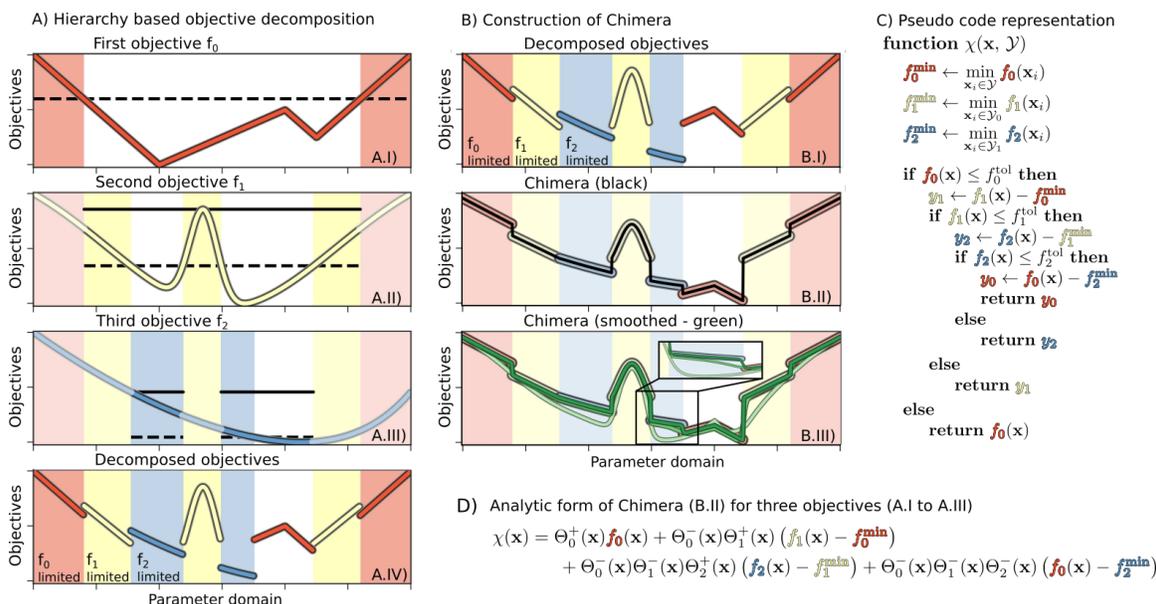


Figure 8.1: Example for the construction of CHIMERA from three one dimensional objective functions. (A) Illustration of the three objective functions, f_0 , f_1 and f_2 , in order of the hierarchy. For constructing CHIMERA, each objective is considered only in the parameter region where higher-level objectives satisfy the tolerances (dashed lines). Solid lines indicate the upper objective bound in the region of interest used as a reference for the tolerance on the considered objective. The objective functions considered in different parameter regions for this example are illustrated in A.IV. (B) The construction of CHIMERA for the considered objective. The discrete variant of CHIMERA (black, panel B.II) is constructed using Eq. 8.2, which was substituted with Eq. 8.6 to generate smooth variants (green, panel B.III) using different smoothing parameter values, where lighter traces correspond to larger parameter values. (C) pseudo code showcasing the conceptual implementation of CHIMERA. (D) Analytic expression for the discrete CHIMERA variant constructed from three objective functions. Reproduced from Ref. [235] with permission from the Royal Society of Chemistry.

Note, that absolute tolerances for individual objectives are computed from the minimum and maximum of this objective only in the subset of the parameter space, $\mathcal{Y}_{k-1} \subset \mathcal{P}$, where the objective one level up the hierarchy satisfies its tolerance criteria (see Fig. 8.1a),

$$f_k^{\text{tol}} = \tilde{f}_k^{\text{tol}} \left[\max_{\mathbf{x}_i \in \mathcal{Y}_{k-1}} f_k(\mathbf{x}_i) - \min_{\mathbf{x}_i \in \mathcal{Y}_{k-1}} f_k(\mathbf{x}_i) \right]. \quad (8.1)$$

We can determine whether a given objective function value is above or below the given tolerance via the Heaviside function Θ ,

$$\Theta \left(f_k^{\text{tol}} - f_k(\mathbf{x}) \right) = \begin{cases} 0 & \text{if } f_k(\mathbf{x}) \geq f_k^{\text{tol}}, \\ 1 & \text{if } f_k(\mathbf{x}) < f_k^{\text{tol}}. \end{cases} \quad (8.2)$$

For the following considerations we introduce the abbreviations

$$\Theta_k^+(\mathbf{x}) = \Theta\left(f_k^{\text{tol}} - f_k(\mathbf{x})\right), \quad (8.3)$$

$$\Theta_k^-(\mathbf{x}) = \Theta\left(f_k(\mathbf{x}) - f_k^{\text{tol}}\right) = 1 - \Theta_k^+(\mathbf{x}). \quad (8.4)$$

Using the Heaviside function to weight the involved objectives, a single ASF can be constructed. This ASF is sensitive only to a single objective in any region of the parameter space (see Fig. 8.1a.iv). However, the assumed values of different objective functions in their respective regions of interest can differ greatly. As such, the value of a lower-level objective might exceed the value of a higher-level objective, as illustrated in Fig. 8.1a.iv. The decomposition of objectives alone would not present a suitable ASF as parameter regions satisfying tolerances on some objectives might be disfavored due to large values of lower-level objectives. To overcome this limitation we propose to shift objectives f_k based on the minimum of f_{k-1} in the parameter regions $\mathcal{Y}_{k-1} \subset \mathcal{P}$ for which f_{k-1} does not satisfy the defined tolerance. We denote the shifting parameters with f_{k-1}^{\min} . CHIMERA $\chi(\mathbf{x})$ is constructed to account for the hierarchy of individual objectives via Eq. 8.5. Following this procedure, the construction and implementation of CHIMERA are illustrated in Fig. 8.1

$$\chi(\mathbf{x}) = f_0(\mathbf{x})\Theta_0^+(\mathbf{x}) + \prod_{k=0}^{n-1} (f_0(\mathbf{x}) - f_{n-1}^{\min}) \Theta_k^-(\mathbf{x}) + \sum_{k=1}^{n-1} (f_k(\mathbf{x}) - f_{k-1}^{\min}) \Theta_k^+(\mathbf{x}) \prod_{m=0}^{k-1} \Theta_m^-(\mathbf{x}). \quad (8.5)$$

Within this formulation of the ASF, and its associated relative tolerances, a single-objective optimization algorithm is motivated to improve on the main objective. The algorithm will also be encouraged to optimize the sub-objectives, from the beginning of the optimization procedure on. Nevertheless, improvements on the sub-objectives will not be realized if they cause degradations in objectives higher up the hierarchy (see Fig. 8.1b.ii). The constructed ASF will be monotonic in proximity to the points in parameter space where CHIMERA transitions from being sensitive to one objective to being sensitive to another objective *if and only if* the two objectives do not locally compete with each other. Detailed explanations on

this property of the constructed ASF are provided in the appendix of Ref. [235]. Identifying the parameter regions where the ASF is monotonic opens up possibilities for interpretations and the potential discovery of fundamental underpinnings.

As the Heaviside function is not continuous, the constructed ASF also contains discontinuities. However, these discontinuities can be alleviated with the logistic function as a smooth alternative to the Heaviside function

$$\theta_{\tau}\left(f_k^{\text{tol}} - f_k(\mathbf{x})\right) = \left[1 + \exp\left(-\frac{f_k^{\text{tol}} - f_k(\mathbf{x})}{\tau}\right)\right]^{-1}, \quad (8.6)$$

where $\tau > 0$ can be interpreted as a smoothing parameter. Note, that the logistic function converges to the Heaviside function in the limit $\lim_{\tau \rightarrow 0^+} \theta_{\tau}(f) = \Theta(f)$. Fig. 8.1b depicts CHIMERA constructed with different values of the smoothing parameter. In general, we observe that small values of τ still retain sharp features in the ASF, although discontinuities are lifted. Large values of τ , however, may cause a deviation in the global minimum of the ASF and in the location of the Pareto-optimal point. The impact of the smoothing parameter on the performance of an optimization run is reported in the appendix of Ref. [235]. We ran PHOENICS on the three one-dimensional objective functions illustrated in Fig. 8.1 and construct CHIMERA with different smoothing parameter values. We find that generally large values of τ result in considerable deviations in the objectives after a given number of optimization iterations, eventually causing the optimization algorithm not to find parameter points yielding objectives within the user-defined tolerances. In contrast, small values of τ (including $\tau \rightarrow 0^+$) cause the optimization algorithm to require slightly more objective function evaluations to find parameter points yielding objectives within the defined tolerances. However, we did not observe any significant differences in the performance for intermediate values of τ . We recommend the use of τ within the $[10^{-4}, 10^{-2}]$ interval. For all the tests performed and reported Sec. 8.3 as well as for the two applications a value of $\tau = 10^{-3}$ was used.

8.3 PERFORMANCE TESTS ON SYNTHETIC BENCHMARKS

The benchmarks presented in this section allow assessing the ability of CHIMERA to find Pareto optimal solutions using single-objective optimization algorithms. We start with a focus on the question of whether CHIMERA locates Pareto optimal points for a given set of hierarchies and tolerances. We then proceed with evaluating the performance and behavior of different single-objective optimization algorithms on CHIMERA. To benchmark the performance of CHIMERA, we consider six different sets of well-established synthetic objective functions. Five of the sets consist of two objectives, while the sixth set contains three objectives. Details on the objective functions are reported in the appendix of Ref. [235]. For all benchmark optimizations reported in this section, we employed the same set of tolerances and constraints on the objectives in the benchmark set, which are reported in the appendix of Ref. [235] as well.

8.3.1 DEVIATIONS OF THE EXPECTED OPTIMUM FROM THE ACTUAL OPTIMUM

The performance of CHIMERA is compared to the behavior of the ASF introduced by Walker *et al.*,¹⁹² which we refer to as c-ASF hereafter due to its constrained approach. Pareto-optimal points were determined from $1,000 \times 1,000$ grids on the parameter spaces. While tolerances on the objectives for CHIMERA can be defined *a priori* without detailed knowledge about the shapes of the objectives, the introduced c-ASF requires absolute constraints on the objectives. For a fair comparison between the two ASFs, we compute constraint values matching the pre-defined tolerances from this grid evaluation. After these initial computations, we emulate an optimization procedure set up as a grid search, which is a common strategy for experimental design.^{188–190} During the optimization, we construct both CHIMERA and c-ASF from obtained observations. We designed the grid from 20×20 equidistant parameter points. From the resulting 400 grid points, we construct 25 different sampling sequences by randomly shuffling the order of grid points. All objective functions are evaluated at parameter points in sequential order. At each iteration in the optimization

procedure, we reconstruct both ASFs and determine their predicted Pareto optimal points. Deviations in the objective values of the predicted and correct Pareto optimal points are used as a measure to determine how well either ASF predicts Pareto optimal objectives. Average deviations between predicted and correct Pareto optimal objectives, relative to the full range of all objectives, are reported in Fig. 8.2a.

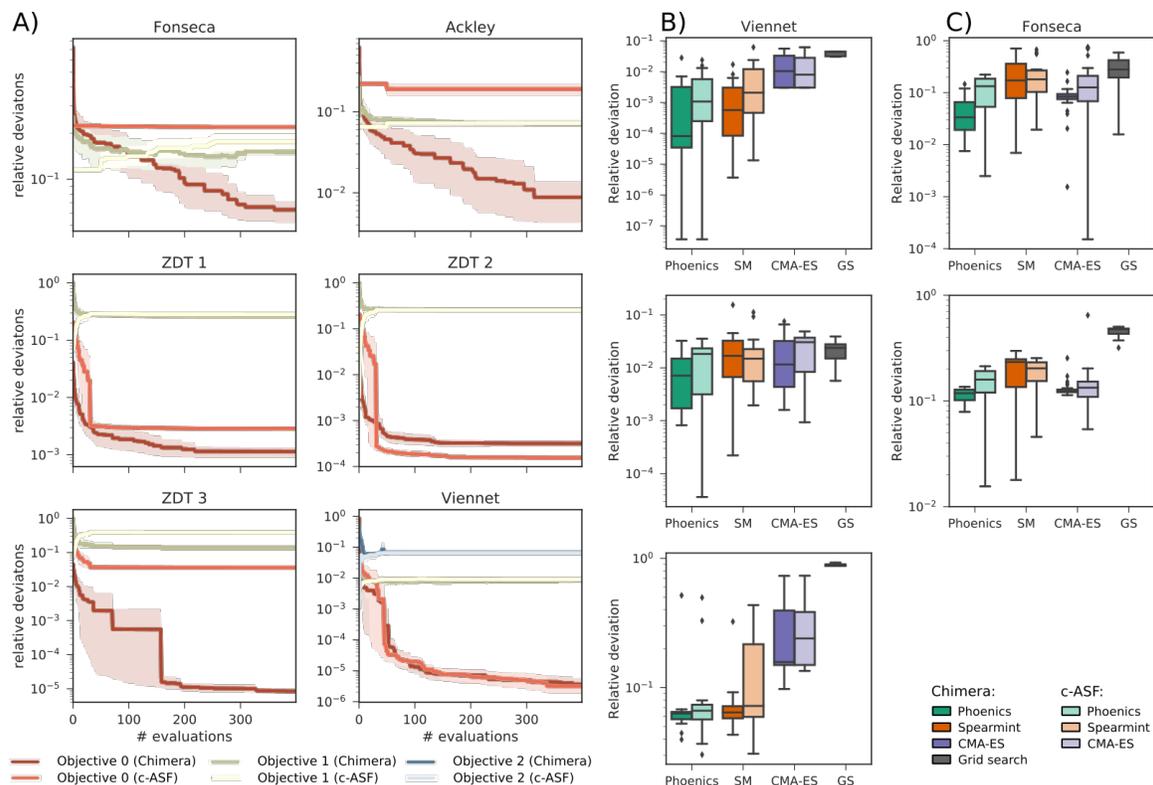


Figure 8.2: (A) Average relative distance from the Pareto-optimal point determined by the applied constraints. We compare the achieved relative distances of CHIMERA and c-ASF. Parameter spaces were searched *via* a grid search (see main text for details). (B) and (C) Average smallest relative deviations between objectives sampled by different optimization algorithms after 100 objective function evaluations averaged over 25 different optimization runs. (B) reports results on the Fonseca benchmark set, and (C) depicts results for the Viennet variant benchmark set. Reproduced from Ref. [235] with permission from the Royal Society of Chemistry.

Based on the benchmark results, we find that the Pareto optimal point predicted by CHIMERA is closer to the true Pareto optimal point for all involved objectives after the full evaluation of the 20×20 grid for four out of the six benchmark sets. With the *Viennet* benchmark set, we find similar performance in both ASFs, and c-ASF predicts Pareto optimal with slightly smaller deviations on the ZDT2 benchmark set. Details on the benchmark sets

are provided in the appendix of Ref. [235]. Besides the prediction accuracy, it is essential to emphasize a major difference between CHIMERA and c-ASF: c-ASF requires detailed knowledge about the individual objective surfaces to set appropriate constraints. The Pareto optimal point can only be determined if reasonable bounds have been defined. Changing the hyperparameters in c-ASF can also significantly influence how individual objectives are balanced. CHIMERA, however, only contains a single hyperparameter, τ , (see Eq. 8.6), which is used for smoothing the constructed χ . From the presented benchmark, we find that CHIMERA shows good performance with the same choice of τ on a diverse set of benchmark functions. We have also illustrated that the performance of an optimization procedure augmented with CHIMERA only weakly depends on the particular choice of τ over several orders of magnitude (see appendix of Ref. [235]).

8.3.2 PERFORMANCE WITH VARIOUS OPTIMIZATION ALGORITHMS

In this section, we report the performance of four single-objective optimization algorithms with both CHIMERA and c-ASF. Specifically, we employ four gradient-free optimization procedures: grid search,^{188–190} covariance matrix adaptation evolution strategy (CMA-ES),^{206,207} SPEARMINT^{237,509} and PHOENICS (see Chapter 6).²³³ The resulting combinations of optimization algorithms and ASFs are applied to the six synthetic benchmark sets and were used to determine how fast the Pareto optimal points can be located. In all optimization runs, we used the same set of constraints and tolerances as discussed in the previous section. The performance of each optimization algorithm augmented with each of the ASFs is quantified by computing the smallest relative deviation in the objectives between all sampled parameter points and the Pareto optimal point. The average smallest achieved relative deviations after a total of 100 evaluations for the Fonseca set and the Viennet set are reported in Fig. 8.2b,c. Note that the performance of the grid search does not depend on the ASF, as decisions about which parameter point to evaluate next are not updated based on prior evaluations. Results on the remaining four benchmark sets are reported in the appendix of Ref. [235]. We find that optimization runs of different opti-

mization algorithms augmented with CHIMERA reach low deviations to the Pareto optimal points after 100 objective set evaluations. When comparing to the deviations in objectives achieved by optimization algorithms augmented with c-ASF, CHIMERA generally seems to lead optimization algorithms closer to the true Pareto optimal objectives. Although the degree of improvement in the deviations of CHIMERA over c-ASF varies across all objectives, we did not observe a case where c-ASF significantly outperforms CHIMERA. These observations hold for the duration of the entire optimization, as reflected by the individual traces reported in the appendix of Ref. [235]. In particular, the fact that the tolerances are defined relative to the observed range of objectives does not appear to be disadvantageous. Indeed, optimization runs with CHIMERA achieve relatively low deviations in all objectives from the beginning of the optimization procedure. Furthermore, we find that optimization algorithms based on Bayesian methods (SPEARMINT and PHOENICS) generally outperform CMA-ES and grid search, although the degree of improvement can vary with the objectives.

8.3.3 BEHAVIOR OF OPTIMIZATION PROCEDURES

In addition to the differences in the performance of CHIMERA and c-ASF with different optimization algorithms, we observe differences in the general behavior of the optimization runs regarding the trade-off between objectives. Traces generated by optimization algorithms augmented with CHIMERA closely follow the user-defined hierarchy in the objectives. As such, improvements on sub-objectives are only realized if superior objectives are not degraded beyond the specified tolerances. Optimization runs generated from procedures augmented with c-ASF do not strictly follow this hierarchy. Instead, we observe cases where c-ASF appears to favor improvements on the sub-objectives even if these improvements cause degradations in superior objectives. An example is given in Fig. 8.3, where optimization traces of grid search and PHOENICS augmented with both ASFs on the ZDT2 benchmark set are presented.

While CHIMERA only allows for improvements on the sub-objective if the main objective is not degraded substantially, c-ASF favors improvements on the sub-objective over improvements on the main objective. This observation, and the fact that this observation

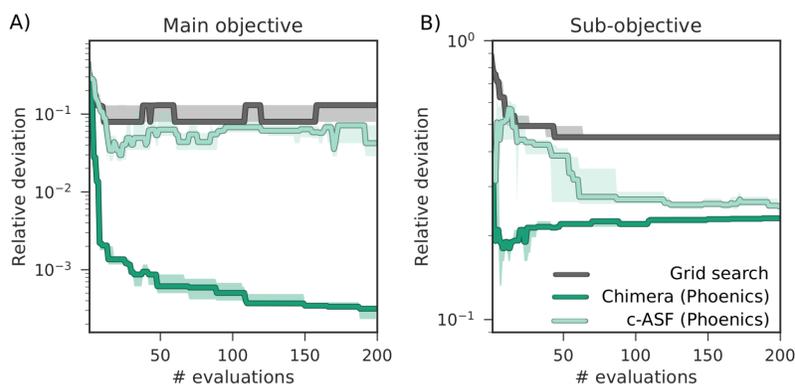


Figure 8.3: Optimization traces representing the smallest relative deviations between sampled objectives and Pareto optimal objectives averaged over 25 individual optimization runs on the ZDT2 benchmark set. (A) shows deviations in the main objective, and (B) displays deviations in the sub-objective. Reproduced from Ref. [235] with permission from the Royal Society of Chemistry.

can only be made for some of the benchmark sets, corroborates with the functional form of c-ASF. Depending on the considered objectives, improvements on sub-objectives can decrease the penalty term such that degradations in the primary objective are allowed. In contrast, CHIMERA strictly enforces the user-defined hierarchy for a wide range of different objective functions, as demonstrated in this benchmark study. In summary, the benchmarks presented in this section illustrate that CHIMERA can identify Pareto optimal points for the provided set of hierarchies and tolerances in the objectives. Moreover, the ASF constructed by CHIMERA enables a variety of optimization algorithms to locate the Pareto optimal point. CHIMERA strictly follows the hierarchy imposed by the user and requires less prior information about the shape of the objectives. Therefore, CHIMERA is well suited for multi-objective optimization tasks where evaluations of the objective functions are costly, satisfying thus the two constraints identified and discussed in Sec. 8.1.

8.4 APPLICATIONS OF CHIMERA

In this section, we demonstrate the applicability and performance of CHIMERA on two different examples: the auto-calibration of a robotic sampling sequence for direct-injection (see Chapter 10), and the inverse-design of a four-pigment excitonic system inspired by

Chapter 3. Both applications involve a larger number of parameters and include three different objectives to be optimized.

8.4.1 AUTO-CALIBRATING AN AUTOMATED EXPERIMENTATION PLATFORM

In this first application, we apply CHIMERA to find optimal parameters for an automated experimental procedure designed for real-time reaction monitoring, as reported in the literature.⁶⁰¹ The procedure is used to characterize chemicals via high-performance liquid chromatography (HPLC). The optimization procedure targets the maximization of the HPLC response while minimizing the amount of sample used in the analysis along with the overall execution time. To benchmark the performance of CHIMERA, experiments were not executed on the robotic hardware, but on a probabilistic model (virtual robot) trained to reproduce the behavior of the real-life experiment, similar to the strategy presented in Chapter 7. The virtual robot is trained on experimental data collected over two distinct autonomous calibration runs orchestrated by the CHEMOS software package (see Chapter 10).⁴¹⁷ During this process, both the HPLC response and the execution times were recorded (see appendix of Ref. [586] for details).

8.4.1.1 CONSTRUCTING A PROBABILISTIC MODEL (VIRTUAL ROBOT)

The virtual robot was set up as a Bayesian neural network (BNN) (see Chapter 5), which was trained to predict HPLC responses and execution times for any possible set of experimental parameters. These parameters were obtained from 1,500 independent experiments conducted fully autonomously without human interaction.^{417,586} For these experiments, the six experimental parameters of the procedure were sampled from a uniform distribution, to ensure unbiased and uncorrelated coverage of the parameter space. For a dense enough sampling of the parameter space, the BNN smoothly interpolates experimental results between two executed experiments. It is important to emphasize that the virtual robot allows querying experimental results for parameters that have not been realized by the actual experimental setup. As such, the virtual robot trained in this work is well suited to inexpen-

sively benchmark algorithms for experiment design. The BNN was trained *via* variational expectation-maximization with respect to the network model parameters. Details on the network architecture, the training procedure and the prediction accuracy on both observed (training set) and unobserved data (test set) are reported in the appendix of Ref. [235]. The probabilistic model is made available on GitHub.⁸⁰

8.4.1.2 EXPERIMENTAL PROCEDURE

The goal of this optimization procedure is to (i) maximize the response of the HPLC, (ii) keep the amount of drawn sample low and (iii) minimize the execution time of the experimental procedure. All results presented in this section were obtained with the PHOENICS optimization algorithm,²³³ and objectives were sampled from the trained virtual robot. PHOENICS was set up with three different sampling strategies, and sequential evaluation of proposed parameter points. We compare the behavior and performance of CHIMERA and c-ASF in two different scenarios, defined by different tolerances and constraints on the individual objectives. By sampling the objective space for 100,000 random uniform parameter points, we can find loose constraints on the objectives such that a parameter point fulfilling all constraints (feasible point) exists. At the same time, such a dense sampling of the parameter space allows us to define a set of objectives that likely cannot be achieved for any set of experimental parameters. As we assume no prior knowledge about the objectives, both scenarios can occur when setting up a new optimization procedure. Based on the 10^5 random uniform evaluations of the probabilistic model, we chose the objective constraints reported in Tab. 8.1 for both scenarios. Tolerances were defined such that they match up with the constraints relative to the entire range of the observed objective function values. Detailed analyses of the influence of each parameter on the objectives, as well as the ranges of the observed objectives, are reported in the appendix of Ref. [235].

Table 8.1: Constraints on the objectives for multi-objective optimization runs on the probabilistic model. Uniform sampling of 100,000 parameter points revealed that loose constraints are achievable by parameter points in a sub-region of the parameter space, while tight constraints cannot be achieved by any parameter point in the parameter space

	Scenario	Response	Sample	Time
Tolerances	Loose	50 %	25 %	50 %
	Tight	20 %	10 %	10 %
Limits	Loose	1250 counts	15 μ l	70 s
	Tight	2000 counts	7.5 μ l	54 s

8.4.1.3 OPTIMIZATION RESULTS

We executed a total of 50 optimization runs with different random seeds and a total of 400 optimization iterations for each set of constraint (loose/tight) and each ASF (CHIMERA/c-ASF). Average traces of the recorded objectives are presented in Fig. 8.4 for loose constraints (A) and tight constraints (B) as defined in Tab. 8.1. When applying loose constraints to the optimization procedure, we observe a similar behavior of CHIMERA and the c-ASF. For both cases, PHOENICS quickly discovers acceptable HPLC responses above the lower constraint, and is then motivated to further minimize the sample volume and the execution time below the specified bounds. We observe a slight trend of CHIMERA, causing PHOENICS to find extensive peak areas after more conducted experiments at the advantage of finding still acceptable peak areas at lower solvent amounts earlier on. This trade-off reflects the hierarchical nature of CHIMERA. With tight constraints, however, we observe a more significant difference between the two optimization strategies. While with both ASFs, PHOENICS finds acceptable peak areas much faster than for loose constraints, CHIMERA appears to help PHOENICS in finding acceptable peak areas in fewer experiments. Moreover, the amount of solvent used in the experiments is lower with CHIMERA from the earliest experiments on and reaches acceptable levels much faster than with c-ASF. However, the upper bound on the execution time is always exceeded, as there is no point in the parameter space for which the peak area is above the chosen lower bound and the execution time below the specified upper bound simultaneously (see appendix of Ref. [235]). CHIMERA, therefore, enables optimization algorithms to rapidly identify parameter points yielding objectives close to the

user specifications. In the scenario where the parameter point does not exist, CHIMERA still leads optimization algorithms to parameter points yielding acceptable objective values based on the provided hierarchy and achieves as many objectives as possible.

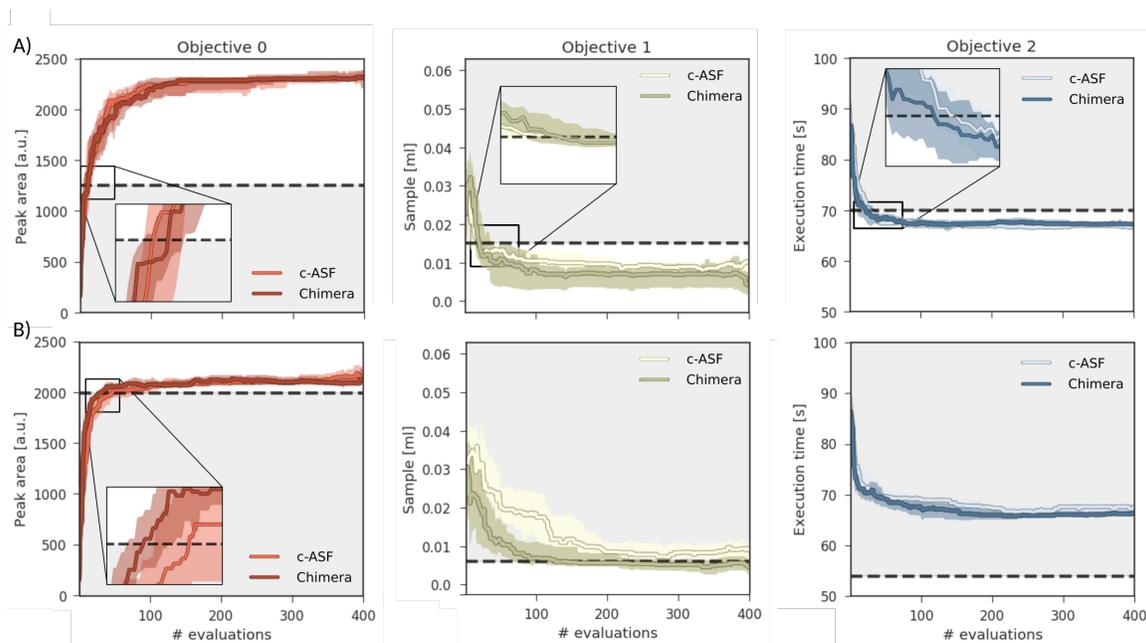


Figure 8.4: Achieved objective function values for multi-objective optimization runs on a virtual robot model obtained with PHOENIX on CHIMERA and c-ASF averaged over 50 individual runs. The goal of the optimization runs is to maximize the HPLC response, minimize the sample volume and minimize the execution time beyond the set bounds, indicated with black dashed lines. Reproduced from Ref. [235] with permission from the Royal Society of Chemistry.

8.4.2 INVERSE-DESIGN OF EXCITONIC SYSTEMS

In this section, we demonstrate the applicability of CHIMERA to inverse-design problems: physical systems are reverse engineered based on desired properties. We focus on the design of a system for excitation energy transfer (EET). EET phenomena have been of great interest in recent years across different fields such as evolutionary biology or solar cell engineering (see Sec. 2.1).^{335,336,382,383} In particular, studies have focused on understanding the relationship between the structure of an excitonic system and its transfer properties fostering the design of novel excitonic devices (see Chapter 3).

8.4.2.1 SYSTEM DEFINITION

The inverse design challenge in this application focuses on an excitonic system consisting of four sites located along the axis, \mathbf{e}_x . Each excitonic site is defined with a position x_i on \mathbf{e}_x , an excited state energy ε_i , a transition dipole with a fixed oscillator strength of $|\boldsymbol{\mu}_i|^2 = 37.1 \text{ D}^2$ and an orientation angle, $\varphi_i = \arccos(\mathbf{e}_i \cdot \mathbf{e}_x)$, with respect to the main axis. As such, the excitonic system is fully characterized by a total of ten parameters: four transition dipole orientations, $\{\varphi_0, \varphi_1, \varphi_2, \varphi_3\}$, three relative excited state energies of the last three sites, $\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$, with respect to the excited state energy of the first site $\varepsilon_0 = 0$ and three relative distances between two consecutive sites, $\{d_1, d_2, d_3\}$, where $d_i = x_i - x_{i-1}$ and $d_0 = 0$. Each of the system parameters was constrained to domains motivated by parameter values for biological light-harvesting complexes.^{129,355–357} Ranges for all parameters are reported in Tab. 8.2.

Table 8.2: Parameters for the excitonic system studied in this application. All parameter ranges are inspired by parameter ranges for biological light-harvesting complexes.

Parameter	size	lower bound	upper bound
Distances d	3	5 Å	40 Å
Energies ε	3	-800 cm^{-1}	800 cm^{-1}
Angles φ	4	0	2π

The goal of the optimization procedure is to design excitonic systems with highly efficient energy transport at low energy gradients across large distances. These three objectives are quantified as follows: assuming the system transfers excitons from the first site to the fourth site, we compute the total transfer distance as $d = d_1 + d_2 + d_3$. Furthermore, we consider the energy gradient between the first and the last site, $\varepsilon = |\varepsilon_3|$. Lastly, we also compute the efficiency, η , of the EET. The transfer efficiency is computed from a full population dynamics calculation in the hierarchical equations of motion (HEOM) approach,^{119,140,141} with the QMASTER software package, version 0.2.^{346,367,602,603} To run a full population dynamics calculation, we construct the Frenkel exciton Hamiltonian^{353,354} for each proposed excitonic system from the system parameters, similar to the procedure outlined in Chapter 3. The

Frenkel exciton Hamiltonian accounts for the excitation energy of each excitonic site and the Coulomb coupling between the sites. While excitation energies are provided as parameters during the optimization, excitonic couplings are computed from the geometry of the system using a point-dipole approximation (see Eq. 8.7).¹¹¹ We denote the unit vector along the spatial displacement of sites i and j with \mathbf{e}_{ij} and the distance between the two sites with d_{ij} . Note that the point-dipole approximation only holds for large distances, which determined the lower bound of 5Å on the distances,

$$V_{ij} = \frac{\mu_i \mu_j}{d_{ij}^3} [\mathbf{e}_i \cdot \mathbf{e}_j - 3(\mathbf{e}_i \cdot \mathbf{e}_{ij})(\mathbf{e}_j \cdot \mathbf{e}_{ij})]. \quad (8.7)$$

The coupling of the excitonic sites, $J(\omega)$, in the system to the surrounding bath are modeled *via* single-peak Drude-Lorentz spectral densities (see Eq. 8.8). For all spectral densities, we chose $\lambda = 35\text{cm}^{-1}$ and $\nu^{-1} = 50\text{fs}$. In all calculations, we use a trapping rate of $\Gamma_{\text{trap}}^{-1} = 1\text{ps}$ and exciton life-times of $\Gamma_{\text{loss}}^{-1} = 0.25\text{ns}$,

$$J(\omega) = 2\lambda \frac{\omega\nu}{\omega^2 + \nu^2}. \quad (8.8)$$

8.4.2.2 OPTIMIZATION PROCEDURE

Calculations of the population dynamics on the described excitonic system are computationally demanding, with execution times ranging from about five to 20 minutes on a single GPU. To accelerate the optimization procedure, we employ PHOENICS, which allows generating multiple excitonic systems per optimization iteration for parallel evaluation. Note, that we extended the sampling procedure in PHOENICS to account for periodicities in the orientation angles by computing periodic distances when constructing the approximation to the objective function from the kernel density distributions. Details on the procedure are provided in the appendix of Ref. [235]. PHOENICS was used with four different sampling strategies, each proposing a different set of parameters in one optimization iteration. For each of the proposed parameter sets, we construct the Frenkel exciton Hamiltonian and

start the population dynamics calculation with *QMaster*. It is important to mention that the execution time of the population dynamics calculation can vary, as it depends on the parameters of the computed system. We, therefore, set up the optimization procedure in an asynchronous feedback-loop, to process results from population dynamics calculations as soon as they are available. In this feedback-loop, a database is used to store system parameters for future evaluations. When a population dynamics calculation completes, a new set of system parameters obtained from the database is submitted for evaluation. Optimization iterations with PHOENICS are triggered right after all three objectives (transfer efficiency, total distance, and energy gradient) have been retrieved from the completed population dynamics calculation. At the end of an optimization iteration, the system parameters in the database are updated with the proposed parameters.

For the task of reverse-engineering an excitonics system, we illustrate the performance of CHIMERA on all possible permutations of hierarchies among all three objectives. For each permutation, we execute a total of 25 individual optimization runs with 400 iterations. All optimization runs aim to design excitonic systems with highly efficient energy transport at low energy gradients across large distances. Note that high transfer efficiencies compete with large distances and low energy gradients. To emphasize the importance of large efficiencies and low energy loss of the transport, we chose to apply a tolerance of 10% on the transfer efficiency, 12.5% on the energy gradient and 40% on the total distance. We find that CHIMERA enables PHOENICS to discover excitonic systems with the desired objectives in all six studied hierarchy permutations. Details about these permutations are provided in the appendix of Ref. [235]. Independently from the order of the objectives in the hierarchy, CHIMERA guides PHOENICS to the parameter region, for which the associated objectives satisfy all tolerances following different sampling paths. We illustrate this in Fig. 8.5, which highlights the objectives sampled for two of the six studied permutations: Permutation 2 (green dots), which (i) maximizes the transfer efficiency, (ii) minimizes the energy gradient and (iii) maximizes the total distance, and permutation 5 (red triangles) which (i) minimizes the energy gradient, (ii) maximizes the transfer efficiency and (iii) maximizes the

total distance. In Fig. 8.5a, we show the points with the most desirable objectives discovered during the optimization runs. Bootstrapped sampling paths leading from the initial (random) points to the best performing points are presented as projections on each of the three planes. Fig. 8.5b to Fig. 8.5d further detail the projected paths by supplementing the individually sampled points for each of the permutations.

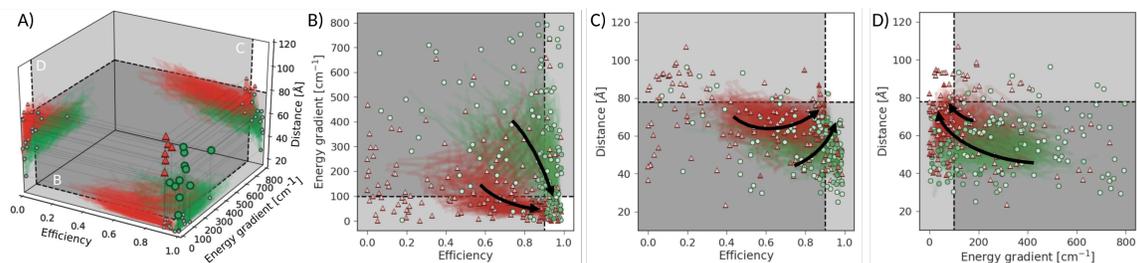


Figure 8.5: Objective function values sampled in optimization runs with two different hierarchies in the objective. Hierarchy order shown in green dots: (i) transfer efficiency, (ii) energy gradient, (iii) total distance. Hierarchy order shown in red triangles: (i) energy gradient, (ii) transfer efficiency, (iii) total distance. (A) Optimal points with respect to all objectives discovered during individual optimizations. Projections illustrate bootstrapped sampling paths leading to the best performing points. (B)-(D) Detailed illustration of projected sample traces. Arrows indicate the general paths taken by the optimization algorithm for the different hierarchy orders. More transparent points have been sampled earlier in the optimization procedure, and more opaque points have been sampled at a later stage. White regions indicate the target values for all considered objectives. Reproduced from Ref. [235] with permission from the Royal Society of Chemistry.

For both permutations presented in Fig. 8.5, CHIMERA successfully leads PHOENICS to the region in objective space where all tolerances are satisfied. However, we observe differences in the sampling paths. While with permutation 2 PHOENICS samples higher transfer efficiencies earlier on in the optimization procedure, the algorithm is biased towards first sampling lower energy gradients with permutation 5. The sampling paths displayed in Fig. 8.5 are in agreement with the order of hierarchies in the objectives for the two permutations. These differences in the samplings paths can be rationalized by the fact that high transfer efficiencies and low energy gradients are competing objectives, *i.e.*, it is not possible to improve on both objectives with the same changes in the parameters. Optimization traces for all permutations averaged over the 25 individual optimizations are reported in the appendix of Ref. [235]. In agreement with previous results on the analytic benchmarks (see Sec. 8.3) and the auto-calibration of an automated experimentation platform (see Sec. 8.4.1) we find that excitonic systems satisfying the primary objective are typically discovered within a few

optimization iterations. Sub-objectives are then easily realized in cases where the first and the second objectives do not compete, *e.g.*, permutation 4, where the first objective is the total distance, and the second objective is the energy gradient. However, if the first and the second objective do compete with each other (*e.g.*, transfer efficiency, and energy gradient in Fig. 8.5) CHIMERA gradually leads to improvements on the second objective without allowing for degradations in the first objective. This behavior is observed across all studied permutations. CHIMERA, therefore, implements the means to realize as many objectives as possible. Based on this observation it can be beneficial to choose the importance hierarchy such that the two most important objectives are expected to not compete with each other to accelerate the optimization process.

8.4.2.3 DERIVING DESIGN CHOICES

In the previous sections, we observed that optimization algorithms strictly follow the implicit objective hierarchy in the ASF constructed by CHIMERA. As such, the excitonic systems sampled during the optimization procedure will achieve objectives in the order of the imposed hierarchy. We now study the excitonic systems sampled during the optimization procedures to retrieve design choices made by the algorithm to achieve the objectives in the imposed hierarchy subsequently. Fig. 8.6 illustrates excitonic systems produced by optimization runs with the following hierarchy: (i) lower the energy gradient, (ii) maximize the transfer efficiency and (iii) increase the total distance covered by the excitonic system. Fig. 8.6a shows the average optimization traces highlighting the portions where only the first objective is reached, the first and second objectives are reached, and all objectives are reached (Fig. 8.6a.i to Fig. 8.6a.iii respectively). Since both low energy gradients and large distances compete with high transport efficiency, only a few parameter points satisfy all three objectives.

Fig. 8.6b illustrates examples of parameters for excitonic systems matching the portions highlighted in Fig 8.6a. The depicted excitonic systems are the earliest encountered sets of parameters in these portions. Arrows indicate both the locations and the orientations of transition dipoles. Associated excited state energies for these sampled systems are pre-

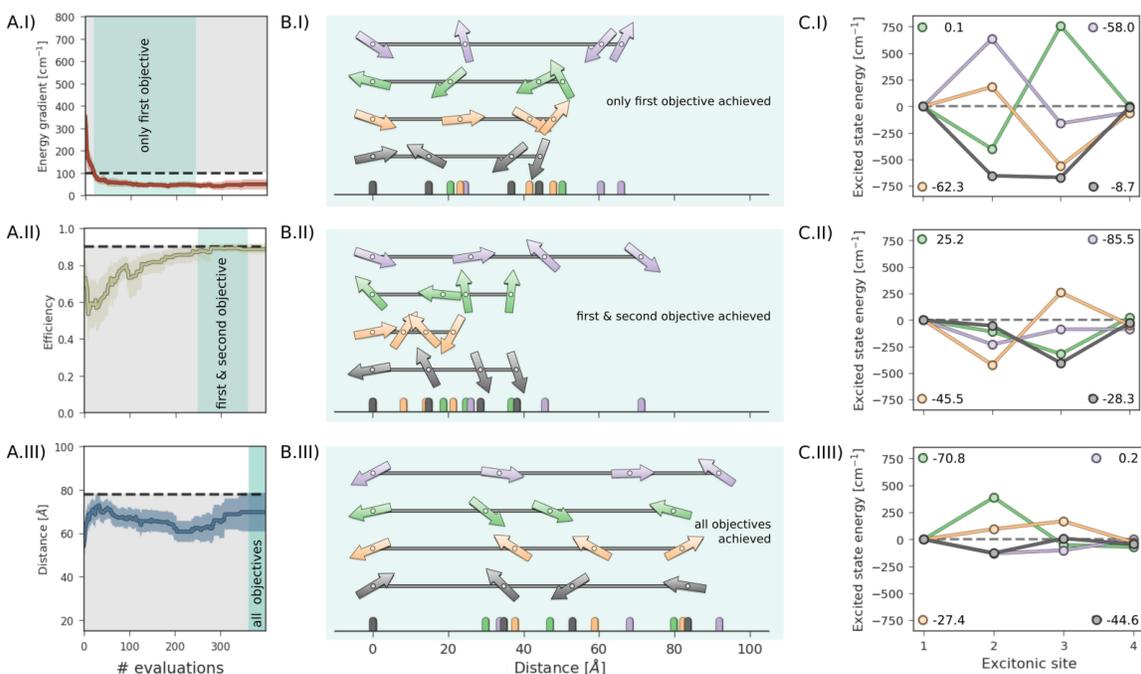


Figure 8.6: Results for the inverse-design of an excitonic system with (i) a low energy gradient, (ii) high transfer efficiency, and (iii) large total distance between the first and the last site. (A) optimization traces averaged over 25 individual optimization runs, and indicate the average required number of designed systems to achieve one, two or all objectives. (B) illustrations of sampled excitonic systems achieving one, two or three objectives. Arrows represent transition dipoles with their location and orientation to the principal axis. (C) Excited state energies of the systems depicted in (B). Overall energy gradients are reported in the legends. Reproduced from Ref. [235] with permission from the Royal Society of Chemistry.

sented in Fig. 8.6c. For the sampled excitonic systems achieving the first objective (low energy gradient, Fig. 8.6.i) we do not observe preferences regarding the distances between excitonic sites, orientations of transition dipoles or excited state energies for all but the last sites. These observations are in agreement with the defined objectives, as the energy gradient is only controlled by the excited state of the last site. To subsequently achieve the second objective (high transport efficiency, Fig. 8.6.ii), we observe a tendency of sampling shorter overall distances and excited state energies which are lower in magnitude. By further constraining the system to maximize the overall distance (Fig. 8.6.iii) transition dipoles are required to align. This sampling behavior provides empirical evidence about the influence of individual system parameters on considered objectives. Overall, we find that CHIMERA is well suited to approach inverse-design challenges and discover physical systems with desired

properties, even if a larger number of parameters determines the properties of the system. In addition, the formulation of CHIMERA in terms of a hierarchy in the objectives allows studying the systems sampled at different stages of the optimization procedure when different objectives are achieved. As demonstrated in the example of designing excitonic systems in Fig. 8.6, general design choices can be evidenced empirically from the sampled systems.

8.5 CONCLUSIONS

In this work, we introduced CHIMERA, a novel ASF for multi-objective optimization tasks associated with experimentation or involved computations. CHIMERA uses concepts of lexicographic methods to combine any n objectives into a single, smooth objective function based on a user-defined hierarchy in the objectives. Additionally, tolerances for acceptable ranges in these objectives can be provided prior to the optimization procedure. CHIMERA strictly follows the imposed hierarchy in the objectives and their associated tolerances. By construction, this approach avoids the degradation of objectives upon improvements on objectives with lower importance along the hierarchy. CHIMERA contains a single hyperparameter, τ , controlling the degree of smoothness of the ASF. However, the performance of CHIMERA appears to be rather insensitive to the value of τ across several orders of magnitude. We nonetheless recommend $\tau = 10^{-3}$ based on our benchmarks. When comparing to the formulation of other *a priori* methods, CHIMERA requires less prior information about the shapes of individual objectives, while providing the flexibility to reach any Pareto optimal point in the Pareto optimal front and keeping the number of objective evaluations to a minimum.

We assessed the performance of CHIMERA on well-established synthetic benchmark sets for multi-objective optimization methods. Our results indicate that CHIMERA is well suited to predict the location of Pareto optimal points following the provided preference information. CHIMERA provides additional flexibility by enabling various single-objective optimization algorithms to efficiently run on top of the constructed ASF. In comparison to the general-purpose constrained ASF suggested by Walker *et al.*¹⁹² we find that CHIMERA enables optimization algorithms to identify Pareto optimal points in fewer objective function evaluations

while requiring less detailed knowledge about the objective surfaces. We further illustrated the capabilities of CHIMERA on two different applications involving up to ten independent parameters: the auto-calibration of a virtual robotic sampling sequence for direct-injection, and an inverse-design problem for excitonic systems. The auto-calibration application revealed that CHIMERA always aims to achieve as many objectives as possible following the provided hierarchy and does not improve on sub-objectives if this would imply degradations of the main objective. This observation is also confirmed with the excitonics application. We found that the imposed hierarchy in the objectives allows deducing design principles from sampled parameters. This possibility can find important applications for molecular and structural design with tailored properties, especially in the context of autonomous experimentation workflows for scientific discovery. Furthermore, it allows to understand the influence of distinct features on the global properties of the system.

With the versatile formulation of CHIMERA, and its low requirements on *a priori* available information, CHIMERA is readily applicable to tasks beyond the scope of the two presented illustrations. We envision CHIMERA to be successfully used in scenarios where slow merit-evaluation processes such as involved computations or experimentation, most notably in chemistry and materials science, present a challenge to other methods. Moreover, CHIMERA enables the use of single-objective optimization algorithms and quickly determines conditions yielding the desired merit. As such, CHIMERA constitutes an important step towards the deployment of self-optimizing reactors and self-driving laboratories in autonomous workflows, as it provides an approach to overcome the identified constraints: (i) objective evaluations involve timely and costly experimentation, and (ii) no prior knowledge about the objective functions is available. In summary, we suggest that researchers in automation and more generally multi-objective optimization test and deploy CHIMERA for Pareto problems when evaluations of the objectives are expensive, and no prior information about the experimental response is available.

Part III

Autonomous experimentation

This page is intentionally left blank.

9

Next-generation experimentation with autonomous platforms

Apart from minor modifications, this chapter was originally published by Elsevier as:

Florian Häse, Loïc M. Roch and Alán Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. *Trends Chem.* **1** (3), 282–291 (2019).

Reproduced from Ref. [81] with permission from Elsevier.

ABSTRACT

The ever-growing demand for advanced functional materials requires to transform established strategies to experimentation and to accelerate the discovery process. State-of-the-art approaches to scientific discovery are inherently time-consuming, resource-demanding, and have arguably reached a plateau. Significant advances could be achieved by rethinking and redesigning the conventional experimentation process with self-driving laboratories. These autonomous experimentation platforms augment automated laboratory equipment with data-driven approaches to enable end-to-end solutions to the experimentation process with little human interaction. To that end, the search for promising materials candidates can be guided by machine learning (ML) models, which hypothesize about the properties and behaviors of novel materials based on empirical evidence collected from previous investigations. This feedback loop is crucial to reduce the number of experiments. Supplying automated platforms with data-driven strategies empowers self-driving laboratories to fully embrace the vision of autonomous experimentation.

9.1 ESTABLISHED AND EMERGING APPROACHES TO EXPERIMENTATION

Many of modern society's challenges spanning science, engineering, and public health demand the discovery of advanced materials.⁶⁰⁴ Recently, functional materials have been designed to address outstanding critical questions in key societal areas, including energy conversion and storage, personalized health-care, and water purification.^{8,11,13,605,606} Yet, materials discovery is an inherently time-consuming and resource-demanding process (see Chapter 1). Typically, it requires up to two decades of basic and applied research for materials technologies to reach the market.⁸⁹ However, a substantial increase of the discovery rate is expected from the combination of automated experimentation with data-driven ML tools into a single platform.⁷ In what follows, we briefly overview state-of-the-art technologies used in modern research facilities before presenting our vision of the next-generation laboratories for an accelerated discovery process.

9.1.1 CONVENTIONAL HIGH-THROUGHPUT EXPERIMENTATION

Modern research laboratories have readily adopted high-throughput (HT) strategies for materials discovery. With HT screening, each material of a pool of candidates is first synthesized and then characterized (see Fig. 9.1a). While experiments are often time-consuming and resource-demanding, the potential of automated experimentation platforms towards accelerated discovery has been realized for decades (see Sec. 2.4).^{303,304,607,608} Automated platforms enable standardization, parallelization, and cost reduction; guarantee reproducibility; and liberate the scientific workforce from repetitive tasks.^{609–611} Early HT screening approaches focused on problems in catalysis^{180,181} and biomedical research.¹⁸² Soon after, applications in materials design followed, *e.g.*, the discovery of methane storage materials¹⁸³ and electrolytes.¹⁹³ The rise of sufficiently accurate computational methods also facilitated the transition from purely experimental HT strategies to computational (or virtual) HT screening and hybrid approaches, where candidate materials are first screened computationally before eventually being verified experimentally.^{612,613} Despite the successes of HT approaches, their

applicability is limited. Combinatorial fabrication of materials yields an exponential explosion of the candidate space, quickly rendering an exhaustive search impractical (see Sec. 2.3). Active approaches, instead, selectively explore this vast space. These strategies reduce the number of experiments yielding discovery to, thus, significantly increase the discovery rate (see Chapters 6 and 7).

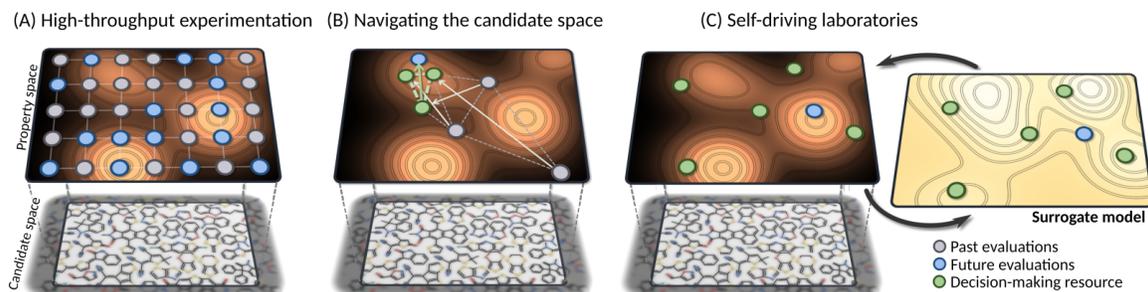


Figure 9.1: Experiment planning strategies. The experiment planning procedure aims to identify experimental conditions for which the resulting properties satisfy desired targets. (A) High-throughput screening iterations consist of scanning a set of a priori selected conditions without refinements. (B) Optimization strategies that plan experiments based on a fixed number of prior iterations. (C) Search strategies that plan experiments based on a surrogate model constructed from all prior experiments. Reproduced from Ref. [81] with permission from Elsevier.

9.1.2 NAVIGATING THE CANDIDATE SPACE WITH SELF-OPTIMIZING REACTORS

Self-optimizing reactors avoid an exhaustive search of the space of potential candidates by reformulating the discovery process as an optimization task. In this formulation, materials are selectively analyzed to identify the material which yields the desired properties. Self-optimizing reactors navigate the application space sequentially, guided by optimization algorithms specifically chosen for the desired application.^{609–611} Recent measurements on candidate materials enable the optimization algorithm to recommend new experiments (see Fig. 9.1b). Early developments of self-optimizing reactors focused on optimizing continuous-flow reactions, for which modified versions of the simplex algorithm proposed experimental conditions.^{198,597} Combinations of factorial experiment designs and gradient-based optimization algorithms have been used for reaction optimization⁵⁹⁹ and identifying the optimal solvent.⁵⁵⁴ The use of grid-based global optimization strategies has also been reported.⁵⁹⁸ The studies mentioned above demonstrate the success of self-optimizing reactors, with software

and hardware specifically designed for the considered application. However, only limited numbers (two to five) of simultaneously changed experimental conditions are reported. Complex applications yield higher-dimensional candidate spaces with more parameters. As the employed optimization algorithms only exploit a fraction of the information acquired from prior experiments, these higher-dimensional candidate spaces are largely inaccessible with only a budgeted number of experiments. Moreover, the application-driven design of these systems is a major obstacle when approaching new problems.

9.1.3 SELF-DRIVING LABORATORIES

Self-driving laboratories accelerate scientific discovery by combining ML for experiment planning with automated experimentation platforms. Parallel to the rapid deployment of automation, steady advances are presented in the field of artificial intelligence (AI), most notably in ML. ML models can identify relations between variables, *e.g.*, reaction conditions and reaction yields, in the absence of physical models (see Sec. 1.3). Trained ML models can use these *learned* relations to speculate about the outcome of a new experiment, enabling autonomous decision making without human intervention. Nevertheless, ML identifies, at best, relevant statistical correlations but does not provide physical insight without the interpretive work of a researcher. ML has been successfully employed to a broad range of chemistry applications, including electronic structure predictions,^{370,375,376} drug design,^{501,510,614} synthesis planning,^{499,615–618} and reaction optimization.⁵⁰⁸ By hypothesizing about the properties of all candidate materials, ML algorithms can recommend promising materials for in-depth analysis (see Fig. 9.1c). The hypothesis is constructed from all previously conducted experiments and is constantly refined with recent measurements. This directed active search of the design space can potentially reduce the number of required experiments for discovery drastically at negligible computational overhead.^{233,235,417,586}

A critical element of the autonomous discovery process is the efficient interplay between the experiment planner and automated platforms in a closed-loop process, where information is continuously shared between the two components (see Fig. 9.2). By leveraging the

strengths of ML and the advantages of automated platforms, the discovery process is thought to be accelerated by an order of magnitude over conventional HT screening approaches.⁷ Typically, five steps are involved in the closed-loop process, which collectively resemble the steps of the scientific method (see Sec. 1.1): (i) The experiment planner defines the experimental strategy to reach the human-defined target. (ii) An approximation to the experimental procedure is constructed via the use of ML algorithms. This approximation is referred to as a surrogate model. With the surrogate model, the experiment planner hypothesizes about the outcomes of all potential experiments. (iii) Experiments are recommended for evaluation if either their outcomes are highly promising or the hypothesis is weak, and the experiment planner is uncertain about its predictions. The first scenario exploits prior knowledge, while the latter explores the application space. (iv) The recommended experiment is then executed and characterized by the automated robotics platform. Based on the performance of the recommended experiment, and on how closely the associated properties agree with the desired target properties, the merit of the recommended experiment is determined quantitatively. (v) The merit serves as feedback for the experiment planner to refine or validate its surrogate model. This feedback mechanism is crucial to closed-loop experimentation as it enables the experiment planner to better hypothesize about the next experiment to be executed.

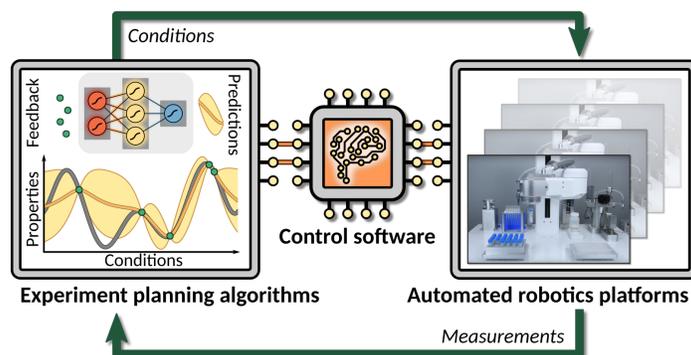


Figure 9.2: The closed-loop approach. The closed-loop approach uses an experiment planning algorithm based on machine learning and a number of automated robotics platforms to execute experiments. Several machine learning methods can be used to construct surrogate models, such as Bayesian neural networks (illustrated earlier), Gaussian processes, or random forests. A control software schedules experiments and directs experimental feedback. Reproduced from Ref. [81] with permission from Elsevier.

One of the first autonomous experimentation systems was reported for the unsupervised growth of carbon nanotubes using random forest (RF) models.^{203,549} More realizations of self-driving laboratories quickly followed, *e.g.*, the production of Bose-Einstein condensates, where experimental conditions are suggested by a Gaussian process (GP).⁵⁴⁶ Other examples include the crystallization of polyoxometalate clusters,⁵⁰⁴ the discovery of NiTi-based shape memory alloys,⁶¹⁹ and the discovery of chemical reactions.^{620,621} While the studies mentioned above present full implementations of self-driving laboratories, other studies focused on algorithmic developments for experiment planning, and define algorithmic goals⁶²² or propose strategies to search for ideal conditions.^{233,235,508} The implementation of a closed-loop requires an efficient and robust control software to orchestrate ML algorithms and automated platforms.^{417,586} Versatile implementations of such software have the potential to operate various automated platforms and execute more complex experiments or different procedures in parallel, as demonstrated in Chapters 10 and 12. An example of a system controlling different, application-specific automated hardware has recently been reported for the optimization of continuous-flow chemical synthesis.⁵⁴⁷

Integrating ML into the experimentation process also allows equipping self-driving laboratories with features that are not immediately required for autonomous experimentation but improve their user-friendliness or enrich their capabilities. It has been demonstrated that self-driving laboratories can be supplied with simple natural language processing (NLP) for convenient communication *via* common messaging services.^{417,586,587} Other ways of interaction have been reported in the form of graphical user interfaces (GUIs)⁵⁴⁷ and web interfaces.⁵⁰³ Self-driving laboratories also provide the opportunity to distribute the experimentation process across several laboratories at different locations by physically disconnecting some of its components. This idea of remote experimentation has been introduced at several levels of partitioning. In early studies, researchers were physically separated from the laboratory equipment.⁶²³ Further developments also disconnected the control software from the laboratory equipment^{503,587} and the experiment planner.^{417,586}

9.2 IMMEDIATE CHALLENGES TO THE MASSIVE DEPLOYMENT OF SELF-DRIVING LABORATORIES

Self-driving laboratories are emerging as the next-generation approach to scientific discovery. Nevertheless, their deployment is limited due to the significant challenges posed by establishing the required robust connection between automated platforms and experiment planning methods to *close the loop*. As a result, self-driving laboratories reported to date mostly focus on well-defined applications despite the opportunity to reach much broader application scopes. A step towards broader applications has only recently been reported in the context of continuous-flow reaction optimization.⁵⁴⁷ Crucial to the flexibility of this system are modularized reactor components that are dynamically integrated into a single workflow. Other studies report a modularization of the control software to orchestrate multiple computation and robotic components (see Chapter 10).^{417,586} This modularization of self-driving laboratories is an essential step towards expanding their applicability and reducing the obstacles to their deployment. As a matter of fact, individual modules equipped with application program interfaces (APIs) act as black-boxes and control, *e.g.*, ML models, or experimental units. Researchers can modify or extend individual modules, and even add new modules to the self-driving laboratory without interfering with existing ones. More importantly, reusing established modules substantially reduces the obstacles to their development.

The experiment planning module interfaces with ML-based active learning algorithms. Suitable algorithms must be robust to noise and recommend only the most informative experiments. Current self-driving laboratories often implement only one learning strategy with very little justification for this choice. The performance of the self-driving laboratory is thus constrained to the implemented learning strategy. A more powerful approach consists in equipping self-driving laboratories with multiple learning strategies. A high-level decision-maker then chooses the learning strategy, which is deemed to be best performing for the considered application. With the enormous potential to leverage from augmenting automated platforms with AI, the development of a self-driving laboratory should not stop

at the implementation of the closed-loop. The experiments conducted by a self-driving laboratory can provide relevant chemical insights even beyond the studied application, as we illustrated in Chapter 5. The ML community has long realized that ML models are best trained with a broad spectrum of examples.⁶²⁴ This observation has recently been verified for materials discovery,³²⁸ where models improved by learning from unsuccessful experiments. The benefits of information-rich training data demand storage of all experimental data in standardized databases, where experiment planners are trained.^{622,625} As self-driving laboratories control every aspect of the experimental procedure at any time, they provide unique opportunities to generate these standardized databases, and directly exploit them for future experiment planning.

Finally, the widespread adoption of self-driving laboratories into conventional laboratory environments critically depends on their user-friendliness (see Sec. 2.4). Little dependence of the control software on proprietary codes or compilers are as important as intuitive interactions for researchers. Existing or emerging technologies and robotic platforms frequently require some degree of expertise in scripting, programming, and algorithmic thinking. These skills are typically orthogonal to the abilities experimental scientists acquire from conventional career paths. Bridging this disparity calls for the implementation of intuitive interfaces, as can be provided by ML-powered NLP. Researchers can conversationally interact with the self-driving laboratory *via* common messaging platforms, which alleviates the challenges of becoming acquainted with command-line interfaces or traditional GUIs.

The closed-loop experimentation process (see Fig. 9.2) is a requirement to reach autonomy in experimentation with self-driving laboratories. Nonetheless, augmenting automated platforms with ML opens up many opportunities to expand the capabilities of self-driving laboratories to facilitate their deployment. Possible expansions are illustrated in Fig. 9.3. Notably, self-driving laboratories can be supplied with complex statistical tools for the on-line analysis of experimental results. Moreover, the integration of NLP modules into a self-driving laboratory offers an intuitive communication interface for researchers. Finally, as self-driving laboratories control every aspect of the experimentation process at any time,

they provide the unique opportunity to store experimentation information in a standardized format.

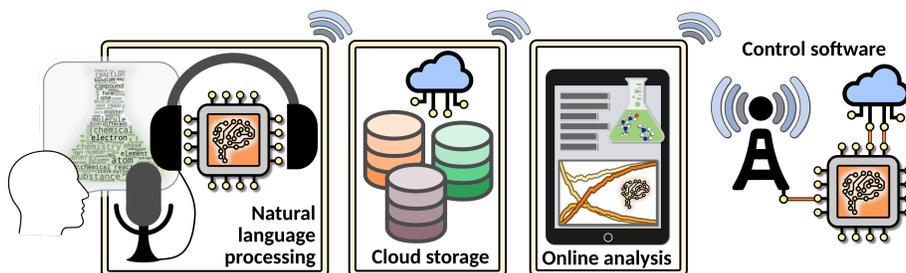


Figure 9.3: Opportunities for enriching features toward user-friendliness. Recent developments in the AI community provide tools which can improve the practicality of a self-driving laboratory. Reproduced from Ref. [81] with permission from Elsevier.

9.3 FUTURE ADVANCES OF SELF-DRIVING LABORATORIES

Supplying automated platforms with recent developments in the rapidly advancing areas of AI, most notably ML, NLP, knowledge management, and intelligent control, enables self-driving laboratories to fully embrace the vision of autonomous experimentation. AI enriches the capabilities of these laboratories in at least two aspects: (i) experiment planning and the orchestration of heterogeneous experimentation platforms, and (ii) the management, processing, and communication of experimental findings (see Fig. 9.4). While advances in the first aspect directly affect the closed-loop process and thus impact the discovery rate, advances in the latter aspect enhance user-friendliness and provide opportunities for researchers to derive scientific concepts based on the conducted experiments.

Self-driving laboratories can simultaneously orchestrate numerous heterogeneous robotics platforms, consisting of, *e.g.*, several individual synthesis and fabrication stations, and various characterization tools, and different experiment planning algorithms in an asynchronous closed-loop. Every individual experiment will be proposed by an experiment planner and executed on the required set of experimentation platforms scheduled by the control software. When supplying the control software with probabilistic ML models, execution times of experiments can be anticipated in advance. Such an intelligent scheduling system can, therefore,

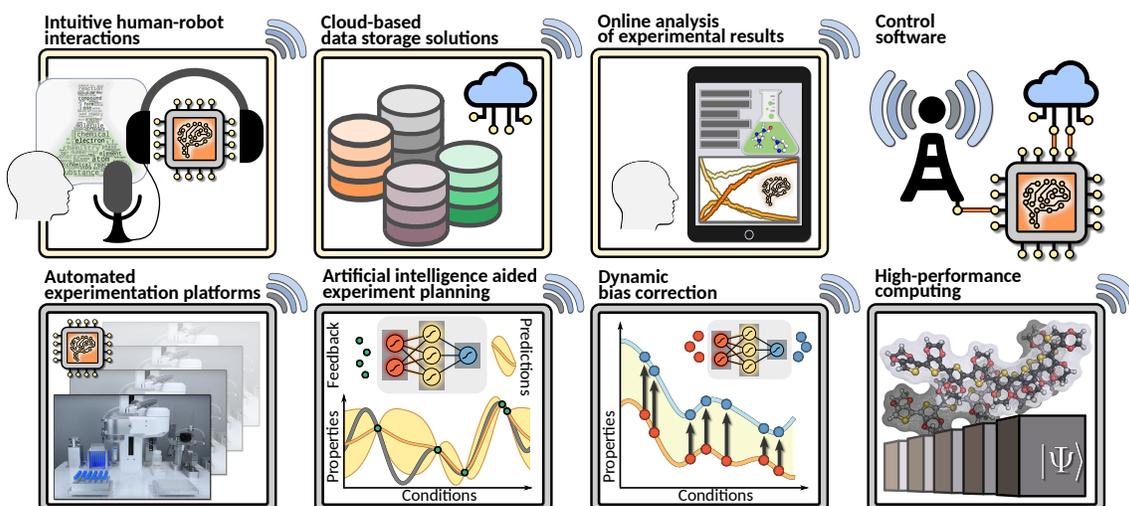


Figure 9.4: The modules involved in next-generation self-driving laboratories. They are empowered by artificial intelligence for experiment planning, device scheduling, data analysis, and researcher communication. A control software (top right) orchestrates individual modules of the self-driving laboratories. Modules highlighted in gray are directly involved in the experimentation process, either to plan new experiments or to execute recommended experiments. Modules highlighted in yellow improve the user-friendliness and practicality of self-driving laboratories and facilitate long-term data storage or communication with researchers. Reproduced from Ref. [81] with permission from Elsevier.

reduce the waiting times for individual experimentation stations and use them to full capacity. With the ability to orchestrate heterogeneous experimentation platforms, self-driving laboratories can be equipped with multiple measurement stations. Some measurements, *e.g.*, the degradation of a material under harsh conditions (see Chapter 12), can be more time-consuming and expensive than simple measurements. Complex closed-loop implementations can combine faster, less precise measurements with slower, more precise measurements to evaluate material properties at several levels of confidence and cost. Intelligent experiment planners can use this heterogeneous feedback to project the outcomes of more demanding evaluations based on the results of cruder evaluations. By balancing the expected outcomes of the more precise evaluations and their associated costs, the experiment planner can choose to discard a recommended material based on the crude property estimates if they are not sufficiently promising. The experiment planner can even actively learn correlations between cheap and expensive evaluation methods. With feedback from multiple evaluators, an additional ML module can learn the bias of the less precise evaluator and construct a correction

model. Similar approaches have been introduced as δ -learning for computing thermochemical properties of small molecules at different levels of theory.⁴⁷¹ Probabilistic bias correction models could inform the experiment planner about the confidence of their correction. The experiment planner can trust the bias-corrected cheap evaluation or verify the prediction with the more precise evaluation. An accurately trained bias correction model can thus be used to avoid expensive experiments entirely.

Augmenting the closed-loop approach with such a bias-correction model also simplifies the integration of computational tools for fast evaluations of material properties. Simple combinations between computational and experimental approaches have long been used in virtual HT screenings. Recently, the benefit of an interleaved workflow between experiments and computations has been reported for the discovery of organic cages.⁶²⁶ Interfacing self-driving laboratories with high-performance computing (HPC) facilities allows for the massive deployment of computational property evaluation strategies in the closed-loop approach. Meanwhile, probabilistic ML models can be trained on-the-fly on experimental results to correct for the bias between computational and experimental evaluations. Self-driving laboratories can collect all information about conducted experiments for long-term storage. With this rich pool of information, data-driven approaches can be leveraged to identify and extract general trends observed across different applications. Researchers can query these trends and conceptualize them to develop insights and formulate scientific concepts. Thus, self-driving laboratories provide the means for interpreting their actions while at the same time detailing evidence for general design choices (see Chapter 7). Interfaces to platforms that clearly visualize these findings will be crucial to improve the human-robot interaction. Identified trends can be used to generate prior expectations on the outcomes of new experimental procedures for different applications. This knowledge transfer is a crucial challenge for self-driving laboratories to accelerate scientific discovery by building up data-driven chemical intuition. While biasing expected experimental outcomes based on past measurements might accelerate the discovery process, scientific discovery can be counter-intuitive and compromise prior expectations. The evolutionary computing com-

munity knows many examples of unbiased algorithms being capable of finding unexpected solutions.⁶²⁷ Self-driving laboratories inherently have the ability to search for novel materials without any bias, which empowers them to make counter-intuitive discoveries. It is, therefore, crucial to not constrain the experiment planner but instead enable it to choose to follow or to disregard the supplied intuition self-consciously.

9.4 THE ROLE OF THE RESEARCHER

The transition from traditional to unsupervised experimentation and the widespread deployment of self-driving laboratories will also impact the role and the responsibilities of researchers, making us redefine what it means to be a scientist. Liberating researchers from arduous, repetitive tasks is not expected to reduce their creativity and, on the contrary, will allow them to focus more on innovative and productive tasks.⁶²⁸ Self-driving laboratories will, therefore, empower researchers to approach more challenging problems of broader scope. Although several processes will be automated within a self-driving laboratory, the researchers will still be required to have a deep scientific understanding of the studied problems. However, specialized technical knowledge, *e.g.*, how to operate particular equipment or execute specific analyses, will become less critical.

9.5 CONCLUDING REMARKS AND FUTURE PERSPECTIVES

The rapid pace of innovations and the many new technologies in ML can take conventional experimentation to the next level. Self-driving laboratories leverage the advantages of ML in combination with automated experimentation platforms for a substantially improved scientific discovery rate. The implications of this development will eventually change the way research is conducted, as the machine-driven inverse-design approach to discovery challenges the conventional Edisonian approach. At the same time, recent advances in NLP and the internet-of-things offer possibilities to implement self-driving laboratories in such a way that they easily integrate into existing laboratory environments and equip today's researchers with conveniently accessible support for experimentation planning and execution. Such a

high level of integration will further accelerate the deployment of self-driving laboratories. Consequently, self-driving laboratories liberate the scientific workforce to focus on more innovative and productive tasks. Researchers are provided with the opportunity to focus on conceptualizing the findings of the many conducted experiments and formulating comprehensible design principles and scientific ideas.

10

ChemOS: an orchestration software to democratize autonomous discovery

Apart from minor modifications, parts of this chapter were originally published by the American Association for the Advancement of Science and the Public Library of Science:

Loïc M. Roch,* Florian Häse,* Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars P. E. Yunker, Jason E. Hein and Alán Aspuru-Guzik. ChemOS: orchestrating autonomous experimentation. *Sci. Robot.* **3** (9), eaaat5559 (2018).

Loïc M. Roch,* Florian Häse,* Christoph Kreisbeck, Teresa Tamayo-Mendoza, Lars P. E. Yunker, Jason E. Hein and Alán Aspuru-Guzik. ChemOS: an orchestration software to democratize autonomous discovery. *PLoS one* **15** (4), e0229862 (2018).

Reprinted with permission from the American Association for the Advancement of Science and the Public Library of Science.

ABSTRACT

The current Edisonian approach to discovery requires up to two decades of fundamental and applied research for materials technologies to reach the market. Such a slow and capital-intensive turnaround calls for disruptive strategies to expedite innovation. Self-driving laboratories have the potential to revolutionize experimentation by empowering automation with artificial intelligence to enable autonomous discovery. However, the lack of adequate software solutions significantly impedes the development of self-driving laboratories. In this chapter, we make progress towards addressing this challenge, and we propose and develop an implementation of CHEMOS, a portable, modular, and versatile software package which supplies the structured layers indispensable for the deployment and operation of self-driving

* These authors contributed equally

laboratories. CHEMOS facilitates the integration of automated equipment and enables the remote control of automated laboratories. CHEMOS can operate at various degrees of autonomy, from fully unsupervised experimentation to actively including inputs and feedbacks from researchers into the experimentation loop. The flexibility of CHEMOS provides a broad range of functionalities, as demonstrated on five proof-of-concept applications, which were executed on different automated equipment, highlighting the versatility of the software package and the advantages of autonomous experimentation workflows.

10.1 AUTONOMOUS APPROACHES TO SCIENTIFIC DISCOVERY

Autonomous laboratories combine automated robotics platforms with data-driven algorithms. This combination optimizes and experimentation process to reach a human-defined target in full autonomy. In autonomous laboratories, data-driven strategies leveraging emerging machine learning (ML) technologies design and suggest experiments, which are validated on the automated robotics platforms. The ML algorithms process the outcome of this validation to refine experimental strategies and, subsequently, to better hypothesize about the next experiment to perform (see Fig. 10.1). This so-called closed-loop approach (see Sec. 9.1.3) contrasts with automated laboratories, where the researchers define the experimental strategies beforehand. Hence, autonomous laboratories have the potential to modernize conventional trial-and-error approaches, thus expediting the discovery of molecules and materials.⁷ Additionally, they can attenuate the clash between expected time to discovery and cost of human-driven experimentation, which has arguably reached a plateau in terms of efficiency gains.⁷

To increase experimental throughput, modern research laboratories have adopted fully automated solutions (see Sec. 2.4), which lead to more efficient optimization of processing conditions to, for instance, screen pharmaceutically active ingredients⁵⁸⁹ or explore wide-bandgap perovskites.³²⁶ However, to capture relevant phenomena eventually yielding discovery, automated platforms require an exhaustive and enduring exploration of the complex and high-dimensional application space.²⁰³ Augmenting systematic search-based automation

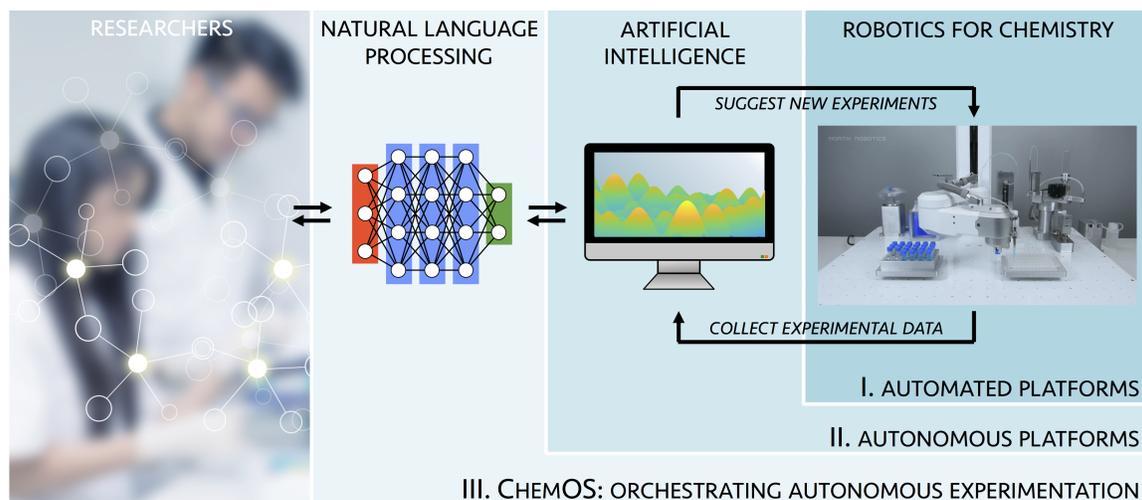


Figure 10.1: CHEMOS a versatile software package orchestrating autonomous experimentation. The set of instructions sent by researchers to CHEMOS is interpreted by a natural language processing module. Data-driven algorithms suggest new experiments to be evaluated on the automated robotics platforms. The experimental outcomes are collected and used to refine the data-driven model of the experiment at hand, in a closed-loop approach. Reproduced from Ref. [417] with permission from the American Association for the Advancement of Science.

with artificial intelligence (AI) may further increase the discovery rate by carrying out an optimal set of experiments based on previous observations.²³³ Recently, several groups have reported promising benefits of autonomous approaches on applications ranging from materials science^{203,504} to organic synthesis.⁵⁰⁸ Although autonomous laboratories are slowly emerging, they are still isolated cases. This development can be partially rationalized by the challenges that the transition from automated to autonomous platforms imposes on engineering software solutions. Suitable software packages are required to (i) orchestrate and synchronize heterogeneous automated instrumentation, (ii) interact with state-of-the-art ML approaches, and (iii) facilitate the communication between researchers and robotics.

In this chapter, we present CHEMOS,^{417,586} a software package designed to enable autonomous experimentation to democratize chemical and materials laboratories. We engineered CHEMOS to be versatile, flexible, and modular. It contains the essential layers for operating autonomous laboratories – administering data collection, experimental procedures, and associated robotics equipment. It also supports the remote control of equipment, which enables CHEMOS to operate across different laboratories. Parallelization techniques

introduced *via* various software design patterns minimize the computational overhead of CHEMOS, allowing multiple experiments to be executed on several robots simultaneously. Consequently, CHEMOS can fully exploit the capacities of both robotics and computational resources. Although CHEMOS consists of six modules, we detail hereafter the three essential ones.

The first critical component of CHEMOS is the learning module, which is the central piece to continuously propose parameters of new experiments to be executed on robotics platforms. The outcomes of previous experiments guide the process for creating new experiments. Once a suggested experiment is evaluated, CHEMOS collects the results from the analytics, *e.g.*, high-performance liquid chromatography (HPLC), nuclear magnetic resonance, and others, and refines its learning procedures on-the-fly to suggest more informative experiments in the next cycle. This closed-loop approach within CHEMOS was validated on two distinct applications:⁵⁸⁶ an in-house developed robot to learn color spaces, pH, densities, and a robotics solution for direct-inject sampling intended for real-time reaction monitoring. To date, four learning procedures are supported, with PHOENICS being the default search strategy (see Chapter 6).

The second key component is the communication module, which facilitates communication between CHEMOS and the researchers through common social media platforms: Twitter, Gmail, and Slack. This interaction is essential,⁶²⁹ particularly in processes requiring human approval, such as those in the food and cosmetic industries. Also, this interaction combines human intuition with data-driven tools in the closed-loop approach to leverage their respective strengths. Hence, the researchers can send commands and instructions at any point in the course of the CHEMOS cycle, including requests for new experiments, status updates, or feedback in plain text messages. A natural language processing (NLP) module within CHEMOS parses the received instructions, simplifying communication while increasing flexibility in the syntax of the messages (see Fig. 10.1). The level of flexibility is such that CHEMOS can be tuned to different degrees of communication ranging from fully autonomous experimentation, as demonstrated on the autocalibration of a robotics solution

for direct injection, to actively integrating human feedback into the closed-loop.

The third key component is the orchestration module, which allows for asynchronous and geographically distributed autonomous laboratories.⁵⁸⁶ Such setups have the potential to improve the efficiency of experimentation by parallelizing experimental procedures and also to reduce costs by taking advantage of robotics platforms acquired and developed by the research groups involved. By sharing complementary fields of expertise, discovery and innovation processes can be accelerated. An example for such distributed workflows was established between Vancouver, Canada, and Cambridge, USA. In this application, CHEMOS calibrated a robotic sampling sequence for direct-inject HPLC analysis. The orchestration capabilities of CHEMOS allow us to run the procedure in full autonomy over several days without human supervision, accumulating a total of 1,100 experiments.

The flexibility and modularity introduced at various levels in the functional architecture enable CHEMOS to power a broad range of autonomous workflows. In that respect, CHEMOS has the potential to follow projects such as Robot Operating System (ROS).⁶³⁰ CHEMOS has an easy-to-install procedure and configuration that facilitate research groups joining the transition from automated to autonomous experimentation. This chapter is a call to arms to the community. We invite all interested experts in robotics, automation, chemistry, engineering, and AI to join us in this quest to accelerate innovation, the critical component of scientific discovery.

10.2 OVERVIEW OF AUTOMATION SOFTWARE IN CHEMISTRY

The industrial sector pioneered automation in search of intensifying chemical and pharmaceutical processes to increase productivity and improve quality (see Sec. 2.4).^{303–309,631} During the last decades, several groups have demonstrated the benefits of automated systems on a variety of chemistries,^{198,303,319–327,601,632,633} and a few laboratory automation software packages have been developed.^{634–637} One step further, the integration of design of experiments (DoE)^{188–190} to automation emerged as a first strategic approach to experimentation.^{191,638} The first closed-loop approach for adaptive experimentation consisted of grid-based surveys,

which identified the most promising starting points for subsequent local optimization *via* different flavors of the simplex algorithm. These hybrid approaches were applied to the optimization of chemical reactions.^{304,607,639,640} However, parameters evaluated from a grid are correlated, which might result in the omission of important features. Grid-based methods also require a substantial number of evaluations to capture relevant phenomena leading, eventually, to discovery. As such, to further streamline the discovery process, the next-generation of autonomous laboratories augments automation with data-driven strategies. By refining experimental procedures based on the most recent observations, data-driven approaches ensure to carry out an optimal set of experiments, avoiding the exhaustive and enduring exploration of the complex and high-dimensional application space (see Chapters 6 and 7).

Several groups have already demonstrated the use of autonomous approaches encapsulating modern AI algorithms to a range of applications. One of the first examples of a self-driving laboratory was reported by Maruyama and coworkers on the synthesis of carbon nanotubes.²⁰³ Similar approaches were used to produce Bose-Einstein condensates (BECs),⁵⁴⁶ optimize organic synthesis reactions,^{508,621} discover multicomponent NiTi-based shape memory alloys,⁶¹⁹ design quantum optics experiments to achieve desired photonic quantum state,⁶⁴¹ and synthesize and crystallize polyoxometalate clusters.⁵⁰⁴ The latter example reports the advantages of such a procedure in terms of accuracy and coverage of the crystallization space by comparing it with human-based and random search approaches. Although the self-driving laboratories appear to be on track to revolutionize the traditional Edisonian approach⁶⁴² to experimentation, they require versatile software to be engineered. This requirement often imposes constraints on the development of such autonomous facilities, preventing their full exploitation. In the examples mentioned above, as well as in other applications reported in the past,^{632,643} in-house developed software packages tied to the hardware and the scientific procedure were used to operate the experimental equipment, parse information, and learn from it. Undoubtedly, distinct scientific challenges require tailored self-driving laboratories. However, a substantial number of processes are common

across them or can be abstracted. Significant advances in high-level programming languages, and computer science have enabled the design of a flexible software package that addresses the specific needs of autonomous laboratories. As such, it becomes possible to provide the scientific community with a structured software package, consisting of fundamental layers common to any self-driving laboratory. These layers require database management, experiment scheduling, collection of experimental feedback, interactions with different learning strategies, and recommendations of experimental conditions for the robotics platform. In what follows, we present CHEMOS, a software package that fulfills these requirements and enables the remote control of robotics. CHEMOS has the potential to increase the discovery rate across chemistry and materials science and also catalyze the realization of prototype self-driving laboratories.

CHEMOS coordinates the overall computational and experimental workflow, monitors experiments, administrates data collection, data storage, and details about the configurations of the available automated laboratory equipment, potentially distributed across different physical laboratories. Among the crucial components of CHEMOS are the data-driven experiment planning strategies encapsulated in the learning module. Although CHEMOS shares the vision of a fully autonomous discovery platform, it allows for efficient interactions between researchers, ML tools, and robotic hardware. To this end, CHEMOS provides various intuitive interfaces for interactions, ranging from simple data exchange via well-defined protocols to NLP. We demonstrate the performance of CHEMOS on four applications, each highlighting different aspects of its implementation. Our findings show the ability of CHEMOS to successfully run at various levels of autonomy, from fully unsupervised experimentation to actively including the researchers in the closed-loop approach to discovery. Also, we confirm the ability of the ML approaches to learn experimental procedures on-the-fly to reach human-defined targets in a minimal number of evaluations on high-dimensional spaces, without prior knowledge. For all tested applications, the same CHEMOS core is deployed on several different Unix-like operating systems to operate different (potentially remote) robotic and characterization hardware, with different state-of-the-art experiment

planning tools, demonstrating the flexibility and modularity of the presented software package. CHEMOS is available for download on GitHub.⁶⁴⁴

10.3 ARCHITECTURE OF CHEMOS

CHEMOS follows a modular architecture composed of a central workflow manager and six independent modules (see Fig. 10.2). Three of the modules are required to enable closed-loop experimentation: (i) ML algorithms for experiment planning, (ii) automation and robotics to execute experiments, and (iii) characterization equipment to assess the performance of the conducted experiment. In addition, CHEMOS provides modules that improve the practicality of the self-driving laboratories: (iv) databases for long-term data storage, (v) intuitive interactions with researchers, and (vi) online results analysis. The modularity of CHEMOS is a crucial element that decouples interdependent tasks. Consequently, CHEMOS can be easily extended by incorporating additional modules, or new features specific to an existing module, without interfering with the established workflow. The modularity of CHEMOS significantly reduces the obstacles to the development and deployment of the self-driving laboratories. Before demonstrating the performance of CHEMOS on applications, we highlight the responsibilities of the individual modules. Detailed descriptions are provided in the appendix of Ref. [586].

DATA-DRIVEN STRATEGIES FOR EXPERIMENT PLANNING. The learning module is key to reach autonomy as it designs experimental campaigns for the scientific procedures requested by the researcher. CHEMOS abstracts experimental procedures to stochastic response surfaces, which describe the merit of proposed generic experimental conditions with respect to user preferences (see Sec. 2.3). The learning module supports various Bayesian ML algorithms for efficient parameter space searches and decision making to recommend conditions for future experiments: PHOENICS,²³³ SMAC,^{241–243} SPEARMINT,^{237,509} and random search.^{184,505}

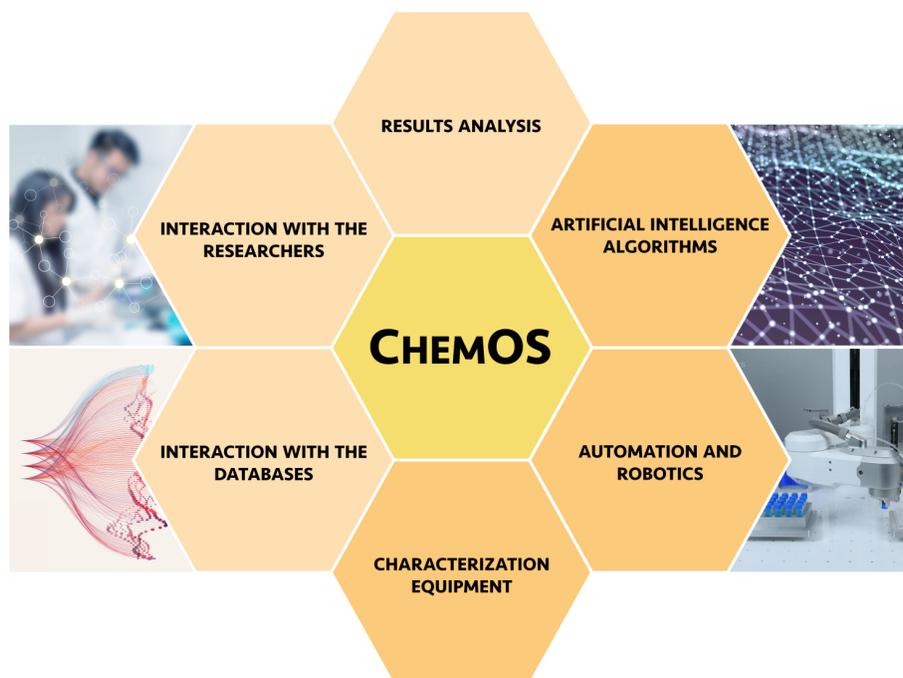


Figure 10.2: Representation of the modules composing CHEMOS. This scheme highlights the modularity and the independence of the six modules, which are (i) global learning procedures, (ii) automated robotic platforms, (iii) characterization equipment, (iv) databases handling and management, (v) intuitive interfaces for researchers, and (vi) online analysis. The central workflow manager, CHEMOS, is depicted in yellow. The required modules to reach autonomy in the discovery process are presented in dark orange. This figure is reproduced with permission from Ref. [586].

INTERACTION WITH AUTOMATED EXPERIMENTATION HARDWARE AND CHARACTERIZATION EQUIPMENT. The robotics and characterization modules provide the mapping between the abstract response surfaces on which the learning procedures operate to the actual scientific procedures. For new applications, the communication protocol and interaction layers facilitating the integration of new hardware and characterization equipment within the CHEMOS workflow need to be implemented manually. It is important to emphasize that this programming task is the only manual step to the deployment, and cannot be automated as the application program interfaces (APIs) may vary from one automated platform to another. Once integrated, the robotics and characterization modules will automatically relate abstract conditions to hardware-specific parameters. Since the robotics and characterization modules contain specificity about the available hardware, the learning module in CHEMOS can be agnostic about actual scientific procedures and can thus be applied to

different procedures simultaneously.

DATABASES FOR LONG-TERM DATA STORAGE. Long-term storage of experimental data and instructions is facilitated via database management system (DBMS). CHEMOS features connections to multiple DBMS for flexible data storage at negligible computational overhead. Efficient closed-loop experimentation is enabled by storing information in distinct databases, which allows for parallelized reading and writing operations. All databases operate on the first-in, first-out (FIFO) principle. Consequently, new requests are queued chronologically, and they are processed as soon as parameters, and the robotic hardware are available.

INTERACTION WITH RESEARCHERS. Rapid advances in ML provide opportunities to redesign the interaction of researchers with experimentation equipment. Approaches such as graphical user interfaces (GUIs) raise the deployment obstacles as researchers need to acquaint themselves with the interface first. CHEMOS provides more intuitive interfaces for researchers in the form of a chatbot framework powered by a NLP module. This interface favors the interaction between researchers, the learning module, and the robotic hardware. The NLP module processes new requests, filters, and summarizes relevant results and customizes responses based on information specific to the received messages (see Fig. 10.4c). CHEMOS connects the NLP module to several common social media platforms and communication services, including e-mail exchange, private and public messaging via Twitter, and private messaging *via* Slack. As such, CHEMOS can accommodate the individual preferences of a researcher or a research group. Communication *via* multiple channels in one session is also supported.

ONLINE ANALYSIS OF EXPERIMENTAL RESULTS. CHEMOS features an analysis module to process, summarize and visualize the results obtained from measurements. This module enables researchers to quickly perceive the progress of the current experimentation and conceptualize the findings of the self-driving laboratory. CHEMOS supports the generation of time traces, runs statistics on repeated experiments, computes higher-level objectives from

lower-level experimental observations, and visualizes the search space of the experimental procedure. Researchers can request a status report at any time in the course of the experimental procedure. Fig. 10.4c illustrates such a status report requested *via* Slack.

ORCHESTRATING EXPERIMENTAL PROCEDURES WITH CHEMOS. The modular implementation and the decoupling of individual tasks of the closed-loop experimentation process enable CHEMOS to orchestrate multiple experimental procedures simultaneously at negligible computational overhead. Individual experimentation sessions can be implemented by providing detailed information and feature selections for each of the modules in a single configuration file. The configuration file includes the researchers' choices of learning procedures and communication channels. Furthermore, it informs CHEMOS about the available automated experimentation platforms. Then, CHEMOS automatically maps experimental procedures to the available hardware. Deploying CHEMOS to different applications only requires to modify the configuration file. This flexibility allows for an accelerated and simplified deployment of the self-driving laboratories and empowers CHEMOS to orchestrate several experiments for different applications.

10.4 MATERIALS AND METHODS

10.4.1 COLOR, PH, DENSITY AND DRINK EXPERIMENTS

The robot was assembled on a frame of aluminum rods and 3D-printed pieces.[†] The pumping system was built with a Raspberry Pi 3 Model B as a microcontroller, a power supply, and a set of eight DC peristaltic pumps of 12V, each connected to a 5V relay. Within this arrangement, the relays were controlled by the Python library GPIO. We used SCP as the communication protocol between the Raspberry Pi and CHEMOS, which was deployed on a Mint environment (Ubuntu-based operating system). The low-level programs controlling the pumps and the RGB sensor, the pH meter, and the scale were written in Python. The measurements obtained by the pH probe (pH probe and circuit by Atlas Scientific), RGB

[†] The 3D model and .stl files were downloaded from GitHub [<https://github.com/ytham/barmixvah>, author: Yu Jiang Tham], in July 2017

sensor (EZO-RGB by Atlas Scientific.), and the scale (Mettler Toledo MS303S) were read on a PC by a serial port with the Python library *serial*, *pylibftdi*, *serial_device2*, respectively. The synchronization between the PC and CHEMOS was performed via Dropbox.

The learning algorithms generate five values on the $[0, 1]$ interval. CHEMOS formats these five values into an eight-dimensional array, according to the specific experimental layout used. Once transmitted to the robot, this eight-value array is re-scaled based on a previous calibration of each pump. Then, the vector is scaled to the total volume to be drawn, *i.e.*, the sum of the five initial solutions. Finally, the RGB sensor measures the color of the solution. CHEMOS receives only the maximum norm distance between the target normalized RGB vector and the measure RGB (also normalized). The latter is minimized during the CHEMOS run. Five initial solutions with the following normalized RGB codes are provided: yellow (RGB = $[0.40, 0.41, 0.19]$), red (RGB = $[0.70, 0.17, 0.13]$), orange (RGB = $[0.57, 0.26, 0.17]$), blue (RGB = $[0.06, 0.36, 0.58]$), and green (RGB = $[0.16, 0.56, 0.28]$). The provided green was chosen as the target color for this procedure. Through the robotics module, CHEMOS automatically maps the proposed experimental parameters to the order of the pumps on the robotic hardware. Importantly, CHEMOS was not constrained to the number of solutions to use. For example, green solutions can be produced by choosing only the provided green mixture or mixing the separate initial yellow and blue solutions. Identical setups were used for the pH and density experiments. CHEMOS communicates with the researcher using Twitter and Gmail. It processes messages through the NLP module, which classifies incoming messages as request or feedback. A set of digital LED strips regulated with the I2C protocol and with the Python library *fcntl* was used to inform the researchers upon completion of a process.

10.4.2 AUTOCALIBRATION EXPERIMENT

The robotic setup was built on a North Robotics N9 platform, paired with an Agilent 1260 Infinite series HPLC. The sample used was 10 mM 1,3,5-trimethoxybenzene in MeCN. The sampling sequence involves first the drawing sample through a needle, and then a sample

loop installed into a Rheyodyne 2-way 6-port selection valve. The valve is then switched, and the diluent solvent is pushed through the sample loop and an in-line mixer to a second loop in another 2-way 6-port selection valve. This second valve is then switched to be in line with the HPLC pump and column, and the HPLC acquisition is triggered. A full cycle of operation involves retrieving a set of parameters from CHEMOS, executing a sampling sequence, rinsing the sampling needle and the push line, retrieval of the integration from the HPLC trace, and return of the integrated value to CHEMOS. Dropbox was used as the communication protocol. The control software for the robot was written in Python, which enabled the variation of several parameters of the sampling sequence. The parameters from CHEMOS are passed to a sampling sequence function, which incorporates those parameters into a sequence of arm movements, pump manipulations, and valve switches. The commands in that sequence are function calls of Python class instances specific to the physical object being manipulated. Pump and valve switch commands are then relayed through the robot to those components. Arm motion commands are calculated and determined dynamically from the target location (sample or rinse vial) before being passed to the robot. The parameter values were obtained as normalized values from CHEMOS, and were scaled to the appropriate ranges. The scalars were user-defined to restrict the values within reasonable ranges accessible by the hardware.

10.5 APPLICATIONS OF CHEMOS

10.5.1 ORCHESTRATION OF STANDARD LABORATORY EQUIPMENT

We suggest a robotic liquid handling platform to produce blends that yield a pre-defined color, pH, and density from a set of starting materials. We demonstrate that CHEMOS is capable of generating such formulations without human supervision and that the produced formulations match the pre-defined goals. The experiments outlined in this section were controlled on and synchronized across three distinct physical platforms: (i) a master platform hosting CHEMOS and the learning procedure, (ii) an automated robotic platform operating the liquid handler (see Fig. 10.3c), and (iii) a characterization platform controlling the

characterization equipment for each of the experimental procedures (RGB sensor, pH meter, and a precision laboratory balance). The communication between the master platform and the characterization platform is supported *via* Dropbox, while the liquid handler and the characterization platform communicate *via* the secure copy protocol (SCP). All communications are supervised and instantiated by CHEMOS from the master platform. The experimental loop is initialized remotely by the researchers. Once the researchers' request is parsed, CHEMOS uses the learning module to recommend the first experiment. The robotics module maps the proposed experiment parameters to the available hardware to execute the experiment. The characterization station measures the properties of interest, which are compared to the target defined by the researcher. Based on this feedback, CHEMOS then recommends promising experiments for future evaluations.

LEARNING THE COLOR SPACE. This procedure generates a target colored solution from a set of five dyed solutions combining three individual dye molecules at different concentrations (see Fig. 10.3). This first use-case demonstrates the closed-loop approach orchestrated by CHEMOS, and highlights the workflow management. CHEMOS was instructed to produce a green solution and could produce the target by mixing red, orange, yellow, green, and blue solutions (see Sec. 10.4 for details). Importantly, CHEMOS was not provided any information about the initial solutions and was not constrained to the number of solutions to use. As such, CHEMOS was free to generate the targeted green from, for example, choosing only the provided green solution, or mixing the yellow and the blue solution.

We ran this experimental procedure with all three implemented experiment planning algorithms (PHOENICS, SMAC, and SPEARMINT) for a total of 25 experiments per session. Fig. 10.3d displays the progress of the experimental procedures for each algorithm. The reported *loss* indicates the maximum distance between the normalized RGB codes of the sampled solution mixtures and the normalized RGB code of the target color. The rapid decrease of the loss validates the capability of CHEMOS to learn the color space in full autonomy. All employed learning algorithms consistently produce mixtures of green color

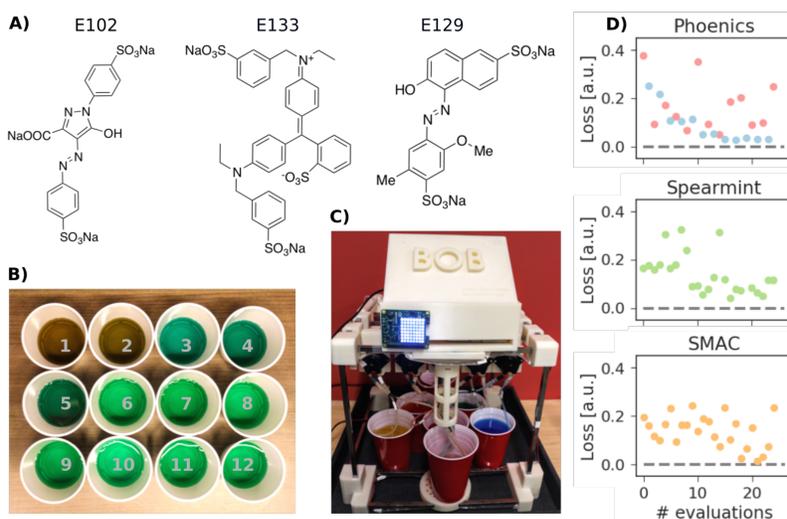


Figure 10.3: (A) The three dyes used in this experiment: E102 (yellow), E129 (red), and E133 (blue). (B) Photograph of the solutions obtained with the 12 exploitation points from PHOENICS (C) Photograph of the in-house built robot. (D) Maximum norm distance (loss) between the achieved normalized RGB color code and the target RGB color code for the 25 experiments. Each panel corresponds to the learning procedure in-use: PHOENICS, SMAC, or SPEARMINT. For the PHOENICS algorithm, red denotes a bias towards exploration, and blue a bias towards exploitation. Note that in exploration mode, PHOENICS samples parameters to gather knowledge where the algorithms has only limited information. In exploitation mode, however, PHOENICS makes the best decision given current information and suggests parameters in the vicinity of the best performing experiment. This figure is reproduced with permission from Ref. [586].

with RGB codes close to the desired target. An example of colors produced by the mixtures sampled from PHOENICS with a bias towards exploitation is displayed in Fig. 10.3b. This simple experiment illustrates the ability of the learning procedures to suggest routes to reach pre-defined targets that deviate from human intuition. Where a human researcher might favor the green starting solution over a mix of multiple starting solutions to produce a green target, CHEMOS suggested a recipe containing 43.6% of the green solution, 19.6% of the yellow solution, 30.0% of the blue solution, and 6.8% the orange solution. This mixture still reproduced the target color indistinguishable to both the human eye and the RGB sensor and presented an unexpected solution to the given task.

LEARNING THE PH SPACE. The goal of this campaign was to generate a solution with a desired pH value using five different starting materials with different pH values. The target pH value is defined by the researcher and set to $\text{pH} = 7.0$ in this example. The five starting materials were prepared with potassium hydrogen phthalate ($\text{pH} = 4.0$, and 5.6),

mixed phosphate (pH = 6.7) and borax (pH = 8.5, and 9.3). CHEMOS used the PHOENICS algorithm to produce the desired target pH within 25 experiments. Fig. 10.4a illustrates how the mixtures proposed by the algorithm approach the target pH of 7.0. Panel 10.4a.i depicts the pH values measured in each experiment, while panel 10.4a.ii shows the best performing composition. No constraints were applied to the number of starting materials to use. We observe a rapid decrease in the deviation of the produced pH value from the desired target for this five-dimensional search space with experiment 10 already yielding a pH of 6.809. The closest experiment to the target is experiment 17, where the pH of the solution was measured to be 7.001. This second illustration showcases the ability of CHEMOS to run in full autonomy. It also highlights the ability of the data-driven algorithm to learn experimental procedures on the fly to reach a human-defined target in a minimal number of evaluations on high-dimensional spaces, without prior knowledge.

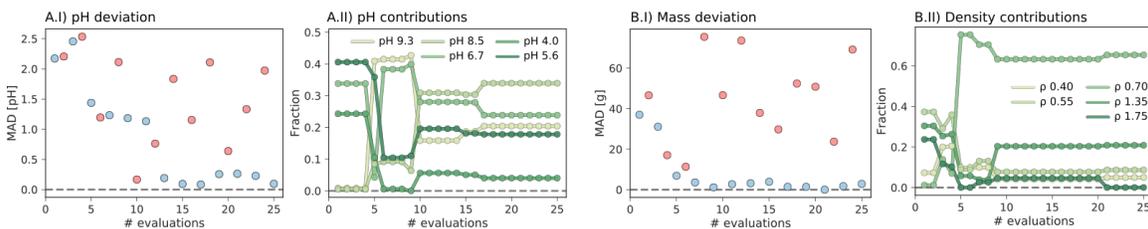


Figure 10.4: Results from the pH (A) and density (B) experiments. Both the loss (i) and the contribution of the starting materials to the produced mixture (ii) are reported. This figure is reproduced with permission from Ref. [586].

LEARNING THE DENSITY SPACE. In this procedure, CHEMOS is provided with five fluids of different densities to produce a blend with a desired target density of 1 g/cm^3 . Like in the previous examples, the densities of the provided fluids (0.40 , 0.55 , 0.70 , 1.35 , and 1.75 g/cm^3) were hidden from CHEMOS. CHEMOS controlled this procedure with the PHOENICS learning algorithm for 25 experiments, following the setup of the pH experiment. Fig. 10.4b reports the deviation of the density of the produced blend from the target density and contribution of each starting material to the blend. Within only nine experiments in this five-dimensional space, PHOENICS reached the targeted density of 1 g/cm^3 . In summary, these three examples demonstrate the capacity of ML technologies to efficiently probe and evolve

on high-dimensional parameter spaces while performing at an optimal number of experiments even without prior knowledge. By not supplying the experiment planning strategies with any prior assumptions, the algorithms are enabled to discover solutions that can be beyond the researchers' expectations. This lack of bias is of utmost importance when targeting scientific discovery. In fact, observations of unbiased ML algorithms finding creative and unexpected solutions have been made recently in the field of evolutionary computation and artificial life.⁶²⁷ Finally, these three applications illustrate the closed-loop approach to experimentation implemented in CHEMOS. They also highlight the seamless integration of different characterization equipment (RGB sensor, pH meter, and precision laboratory balance) into the CHEMOS workflow. The potential of such an approach extends far beyond these applications; we can imagine the potential of CHEMOS when bridged to platform design for drug and material discovery.

10.5.2 AUTONOMOUS CALIBRATION OF A REMOTE ROBOTIC SAMPLING SEQUENCE

In the following, we demonstrate how CHEMOS can be used for remote interactions with distributed automated laboratory systems. We orchestrate an infrastructure capable of unattended sampling. A similar setup was recently used for real-time reaction progress monitoring.⁶⁰¹ The robotics hardware is located in Vancouver, Canada, and controlled by CHEMOS in Cambridge, USA. The procedure involves proposing new calibration parameters, running the experiment, analyzing the experimental results, and updating parameter candidates. The goal of the calibration is to find a set of parameters that maximizes the response of the HPLC, *i.e.*, maximize the amount of drawn sample reaching the detector. The workflow is fully controlled by CHEMOS and does not require human intervention. CHEMOS is set up to communicate with researchers *via* the command line interface for local execution, and *via* Slack for remote execution. The communication between the master platform and the robotic system is enabled *via* Dropbox. Each CHEMOS session is set up for a total of 100 autonomous experiments over a period of about seven hours. We further demonstrate the robustness of CHEMOS on two sessions accumulating a total of 1,400 data points. Details

are provided in the appendix of Ref. [586].

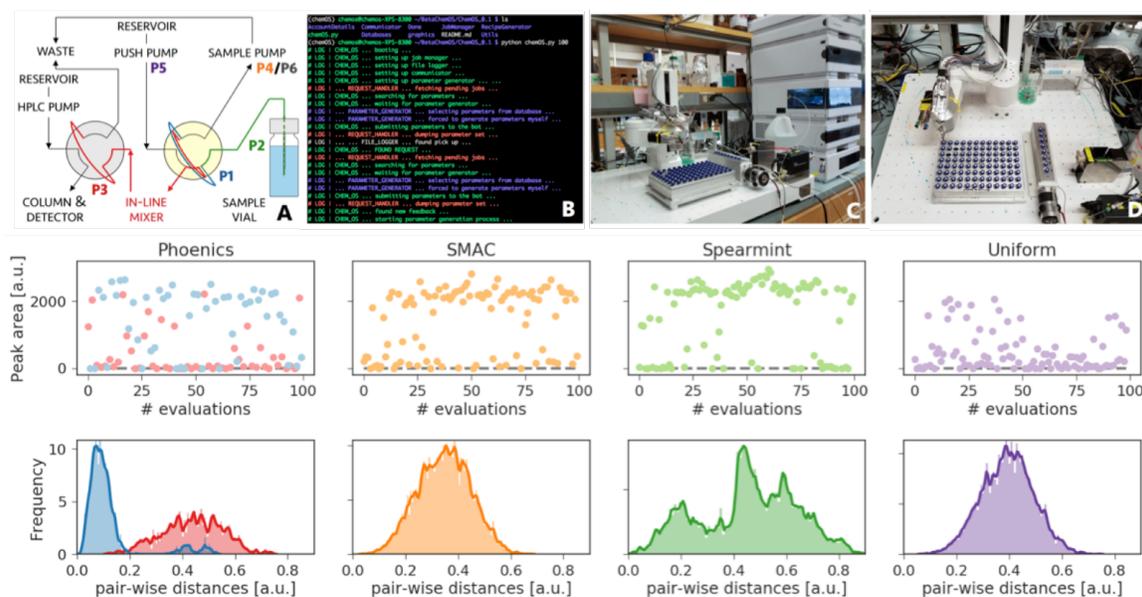


Figure 10.5: (A) Schematic of the flow path for the sampling sequence used with the N9 robotic platform. The six parameters (P1-P6) are color coded to illustrate the effect they have on the sampling sequence. The yellow shade highlights the arm valve, and the grey shade the HPLC valve. (B) Example of logging messages from CHEMOS. (C) Side and (D) top view of the robotic hardware. Lower panels: Results from the autonomous calibration of an HPLC setup maximizing the magnitude of the response. CHEMOS performed autocalibration with four different learning procedures (PHOENICS, SMAC, *spearmint* and Uniform). Upper panels display the achieved peak areas, i.e. magnitudes of response. Lower panels display the distributions of pair-wise distances between sampled parameter points computed with the L2 norm. For the PHOENICS algorithm, red denotes exploration, and blue exploitation. This figure is reproduced with permission from Ref. [586].

The experimental arrangement is illustrated in Fig. 10.5a. A robotic arm (N9 from North Robotics) with an integrated sampling needle is used to draw a sample of 1,3,5-trimethoxybenzene in MeCN from one of the vials in a 96-well tray. The sample is then passed through a sample loop, inline mixer, and injection loop and is finally analyzed by an HPLC. Peak areas of the chromatogram determine the amount of target compound, which was delivered to the HPLC. The entire setup is controlled by six independent parameters determining sample draws, wait times, and push rates. We demonstrate the autonomous execution of the autocalibration by CHEMOS with the four learning procedures: PHOENICS, SMAC, SPEARMINT, and random search. The results of each autocalibration run with 100 individual experiments are depicted in Fig. 10.6. The upper panels display the peak areas achieved during the execution of individual experiments with the different learning proce-

dures. The uniform random searches represent an uninformed exploration of the parameter space as acquired knowledge is not taken into account when proposing new parameter points. We observe that all implemented experiment planning algorithms (PHOENICS, SMAC, and SPEARMINT) quickly find parameter points which yield peak areas larger than those encountered in the random search. The lower panels of Fig. 10.6 depict the parameter distance distributions computed from all possible parameter pairs sampled in each CHEMOS run. While uniform random search appears to yield a unimodal distance distribution, the optimization algorithms show tendencies of favoring broader distance distributions with more parameter points at both smaller and larger distances to other parameter points. We interpret this deviation in the distance distributions as the signature of the optimization algorithms used, in which the acquisition function emphasizes either exploitation of the acquired data (small distances) or exploration of the parameter space (large distances).

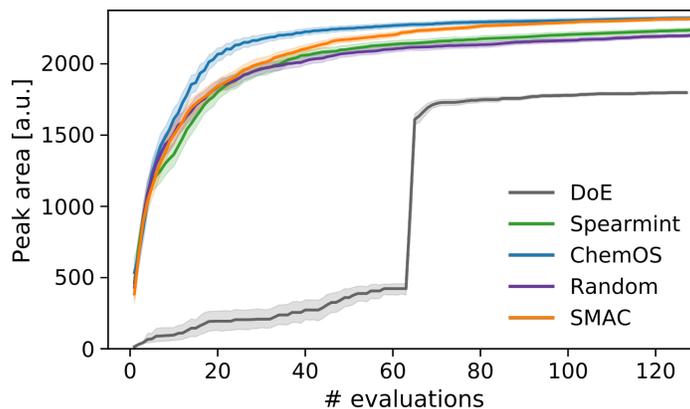


Figure 10.6: Average performance of the experiment planning strategies leveraged by CHEMOS. This figure is reproduced with permission from Ref. [586].

Since repeated executions of individual experiments are time and resource-intensive, we assess the performances of the experiment planning strategies in detail with an emulator constructed to reproduce the experimental surface, following the approach outlined in Chapter 8. Specifically, we train a Bayesian neural network (BNN) to reproduce the response of the HPLC for any given parameter point. Details of the BNN are reported in the appendix of Ref. [586]. The results of the emulator benchmarks averaged over 200 independent executions are illustrated in Fig. 10.6). In addition to the already mentioned experiment planning

strategies, we probed the performance of a DoE approach, where a full factorial grid is used initially to probe the search space and a second, refined full factorial grid is constructed based on the initial responses for an in-depth analysis of a subregion of the search space. We find that DoE is outperformed by the random search strategy, which can be attributed to the high-dimensional nature of the search space. In addition, *spearmint* displays a comparable performance. Only *SMAC* and PHOENICS demonstrate significantly better average performance. With this application, we showed that CHEMOS can control remote facilities to augment automated systems with AI to enable autonomous scientific procedures.

10.6 CONCLUSION

We introduced CHEMOS, a transferable, flexible, and versatile software package. CHEMOS is a framework which supplies all the layers needed to control and orchestrate distributed autonomous laboratories. The five applications reported herein highlight this ability and ease to deploy CHEMOS and to control different applications with a variety of laboratory equipment and automated solutions by only editing the configuration file. The functional design of CHEMOS and its modular structure, in which each module is responsible for fulfilling specific tasks, allows for the global control of complex heterogeneous automation platforms. What is more, the flexible architecture enables to update modules independently and facilitates the addition of new features, which reduces the obstacles to the deployment of self-driving laboratories. CHEMOS includes state-of-the-art data-driven strategies for experiment planning in its learning module. This module is the key component to reach autonomy. Also, to facilitate the researcher-robot interaction, we supplemented CHEMOS with an NLP module in a chatbot framework. This module is based on a neural network and interfaces with several common social media platforms and communication services. It enables the researchers to trigger new experiment from a distance or to send commands and instructions at any point in the course of the CHEMOS cycle. Possible messages include requests for new experiments, status updates, or feedback in plain text messages. Finally, due to parallelization techniques, the negligible computing overhead rising from function

queries, database requests, and data parsing ensures a maximized experimentation throughput. The current version of CHEMOS is the beginning of a comprehensive software package for controlling automated laboratory systems. We believe that CHEMOS has the potential to follow the path paved by the ROS,^{630,645} to become the standard to power self-driving laboratories. The results shown in this manuscript are the first steps towards a global operating system for distributed automated laboratories. Similar to supercomputer centers, which give researchers easy access to computing infrastructure, we envision CHEMOS to democratize autonomous discovery. With access to such laboratories, more research groups could join the evolution of experimentation and could contribute to advances in a broad spectrum of technologies at an accelerated rate.

11

Self-driving laboratory for accelerated discovery of thin-film materials

Apart from minor modifications, this chapter originally appeared as:⁹⁰

Self-driving laboratory for accelerated discovery of thin-films. Benjamin P. MacLeod, Fraser G. L. Parlane, Thomas D. Morrissey, Florian Häse, Loïc M. Roch, Kevan E. Dettelbach, Raphaell Moreira, Lars P. E. Yunker, Michael B. Rooney, Joseph R. Deeth, Veronica Lai, Gordon J. Ng, Henry Situ, Ray H. Zhang, Alán Aspuru-Guzik, Jason E. Hein, Curtis P. Berlinguette. *arXiv preprint* arXiv:1906.05398 (2019).

ABSTRACT

Discovering and optimizing new materials for clean energy applications typically takes over a decade of basic and applied research. Self-driving laboratories that iteratively design, execute, and learn from material science experiments in a fully autonomous loop present an opportunity to streamline the discovery process. In this chapter we report a modular robotic platform driven by a data-driven experiment planning algorithm capable of autonomously optimizing the optical and electronic properties of thin-film materials by modifying the film composition and processing conditions. We demonstrate this platform by using it to maximize the hole mobility of organic hole transport materials for use in perovskite solar cells. An unexpected outcome of this optimization process was the finding that high dopant concentrations can enhance the thermal stability of hole transport films. These results demonstrate the promise of using autonomous laboratories to discover organic and inorganic materials relevant to clean energy technologies.

11.1 INTRODUCTION

Optimizing the properties of thin films is time-intensive because of the vast number of compositional, deposition, and processing parameters available.^{534,646} These parameters are often interdependent and can have a profound effect on the structure and physical properties of the film and any adjacent layers present in a device.⁶⁴⁷ Few computational tools are available for predicting the properties of materials with compositional and structural disorder, and thus the materials discovery process still relies heavily on trial and error. An established method for sampling a large parameter space is high-throughput experimentation (HTE),^{648,649} but it is nearly impossible to sample the full set of combinatorial parameters available for thin films. Parallelized methodologies are also constrained by the experimental techniques that can be used effectively in practice. The overwhelming size of the parameter space for thin film materials motivates the need for both data- and theory-guided algorithms for executing experiments beyond what can be achieved with HTE alone.^{81,556,650} The experimental approach of iterating between automated experimentation and data-driven experiment planning has resulted in early successes in addressing high-dimensional problems in experimental physics,⁶⁵¹ chemistry,⁶⁵² and life-sciences (see Chapter 9).⁶⁵³ This approach is only starting to be implemented in the materials sciences,⁶⁴⁶ as demonstrated by the optimization of carbon nanotube growth,⁵⁰² amorphous alloy compositions,⁵⁴⁴ and inorganic perovskite quantum dot nucleation.⁶⁵⁰ We demonstrate here the optimization of thin films using our platform named ADA, a flexible and modular self-driving laboratory capable of autonomously synthesizing, processing, and characterizing organic thin films. ADA trains itself how to find target parameters without any prior knowledge, enabling iterative experimental designs that maximize the information gain per sample (Fig. 11.1).

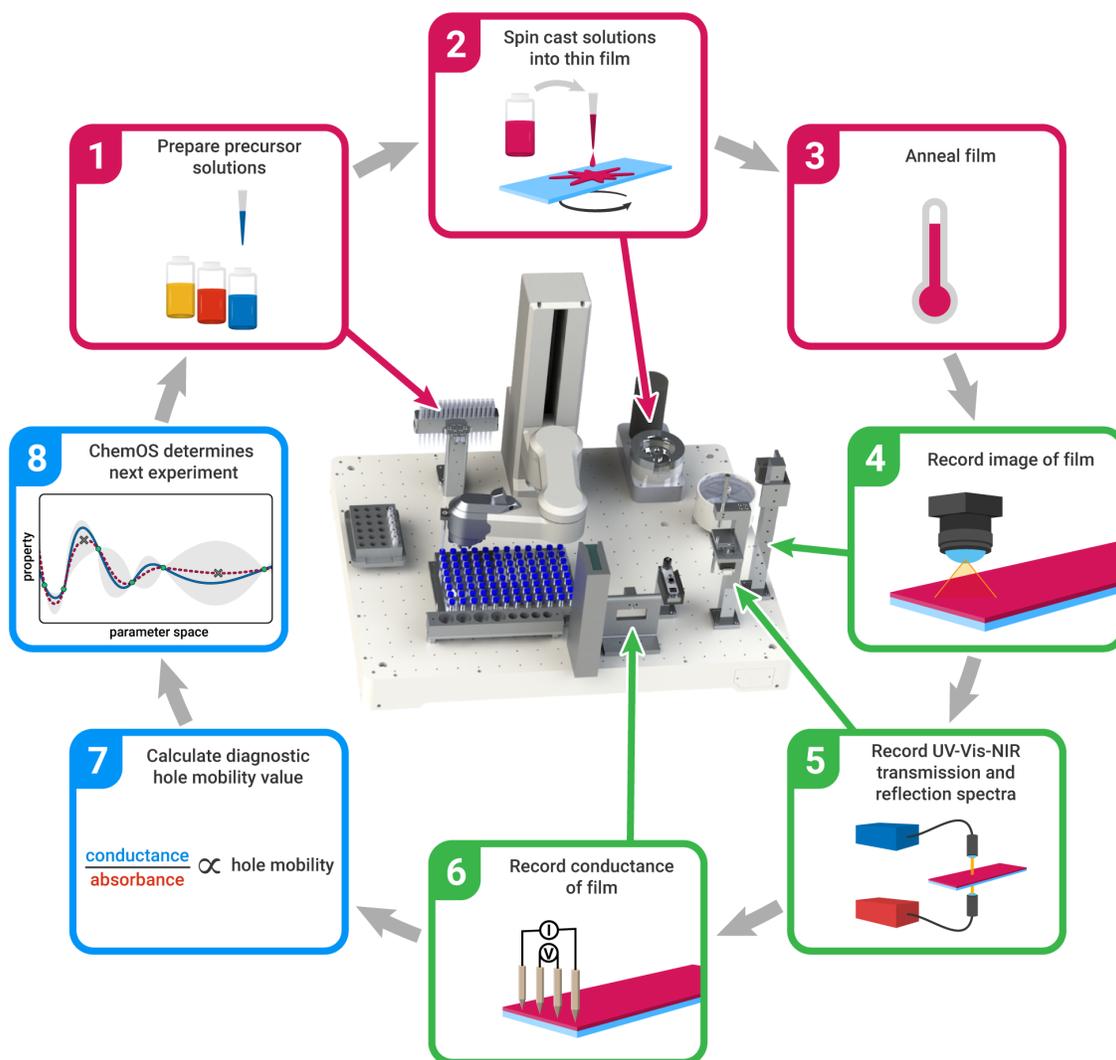


Figure 11.1: ADA implements an autonomous optimization workflow. The self-driving laboratory is based on a modular North Robotics N9 robot (center) and the PHOENICS²³³ optimization algorithm. The robotic arm is equipped with a multi-purpose gripper and a fluid probe connected to a syringe pump (not shown) for performing precision pipetting of solutions. The gripper enables interactions with a variety of objects such as vials, a spin coater, and a vacuum-based substrate handler. The autonomous workflow involves iterative experimentation with the goal of discovering a thin film composition with the highest possible *pseudomobility*. Each iteration of the workflow involves (1) mixing an hole-transport material (HTM)-dopant-additive solution, (2) spin coating the solution onto a substrate, (3) thermally annealing for a variable amount of time, (4) imaging with a visible-light camera, (5) acquiring UV-Vis-NIR spectra in reflection and transmission modes, (6) measuring the I-V curve of the film with a 4-point probe, (7) computing a pseudomobility based on the IV and spectroscopic data, and (8) feeding this pseudomobility into the CHEMOS⁴¹⁷ orchestration software and the PHOENICS optimization algorithm²³³ which then designs the next experiment. Reproduced with permission from Ref. [90].

11.2 RESULTS

As a first step in proving out the methodology, we designed ADA to target organic hole and electron transport layers that are ubiquitous in advanced solar cells,⁶⁵⁴ as well as optoelectronic applications such as organic lasers⁶⁵⁵ and light-emitting diodes.⁶⁵⁶ We demonstrate the capabilities of ADA specifically by optimizing the hole mobility of spiro-OMeTAD, an organic HTM common to perovskite solar cells (PSCs).⁶⁵⁷ The hole mobility of spiro-OMeTAD is critical to PSC performance, but it is highly sensitive to dopants, additives, spin-coating solvents, and post-deposition processing.^{657–661} How each of these factors affect the hole mobility of amorphous spiro-OMeTAD remains difficult to model,^{647,662} and the relevant properties of spiro-OMeTAD are still optimized empirically. This optimization process often takes months to complete and slows the translation of new organic hole and electron transport layers for solar cells and related devices.

ADA autonomously optimizes the hole mobility of spiro-OMeTAD by (i) measuring and mixing solutions of HTMs, dopants, and plasticizers; (ii) depositing solutions as thin films on rigid substrates; (iii) imaging each film to detect morphologies, defects, and impurities; and (iv) characterizing the optical and conductivity properties to produce surrogate hole mobility data. These data are received by CHEMOS,⁴¹⁷ which uses PHOENICS²³³ to design new experiments by actively learning from previously collected measurements (see Chapters 6 and 10). PHOENICS uses a sampling parameter to explicitly bias the experimental design towards exploration or exploitation in an alternating fashion, and has been shown to outperform random and systematic searches.^{233,235,417} The platform aspirates, dispenses, and mixes liquid precursors with the assistance of a syringe pump and a weigh scale. Precursor solutions are spin-cast as thin films on glass substrates, which can then be annealed up to 165 °C using a forced convection annealing system. ADA then characterizes the films using purpose-built systems for darkfield photography, UV-Vis-NIR reflection and transmission spectroscopy, and 4-point probe conductance. The robot also serves as an XYZ sample-positioning stage enabling all characterizations to be performed at multiple positions on the sample, which

we leverage to collect spectroscopy and conductance data at seven spatial positions on each sample. The ability to produce high quality, well-organized datasets while also enabling typically uncontrolled variables (*e.g.*, time between process steps, the height of spin coating dispense nozzle) to become controlled or optimization parameters are compelling features of ADA. Moreover, ADA is operated using flexible, open-source PYTHON software, which facilitates the rapid implementation of new experiments.

We selected HTM hole mobility as our target parameter for optimization, but this parameter typically requires the assembly of multilayer devices to obtain a valid measurement.⁶⁶² Conventional methods are not compatible with the time scales needed for efficient autonomous optimization.⁶⁶³ We, therefore, developed a scheme where we could use 4-point-probe conductivity and UV-Vis-NIR spectroscopy measurements to produce a diagnostic quantity, *pseudomobility*, that is proportional to hole mobility. Pseudomobility is the quotient of the sheet conductance of a thin film and the absorptance of oxidized spiro-OMeTAD in the film. This ratio, which provides a thickness-independent low latency analytical surrogate for hole mobility, became our target optimization objective.

The pseudomobilities of spiro-OMeTAD thin films were optimized by iteratively designing film compositions with variable annealing times and concentrations of dopant. Samples prepared from stock solutions of spiro-OMeTAD and a cobalt(III) dopant (along with a fixed amount of the plasticizer, 4-tert-butylpyridine) were spin-coated onto substrates to yield thin films. Each film was annealed, imaged, and analyzed to determine a pseudomobility value that was relayed to CHEMOS. Fig. 11.2a chronicles how the doping ratios and annealing times were varied during optimization for two independent experimental campaigns. An important outcome is that both campaigns converged on the same global maximum for both doping ratio (~ 0.4 eq.) and annealing time (~ 75 s) to deliver films with the same maximum pseudomobilities. This reproducible endpoint is significant and demonstrates that ADA can successfully navigate a large experimental space.

Fig. 11.2b shows the locations and sequence of the experimentally sampled points in the parameter space. The sampled points can be seen to initially cluster at a local max-

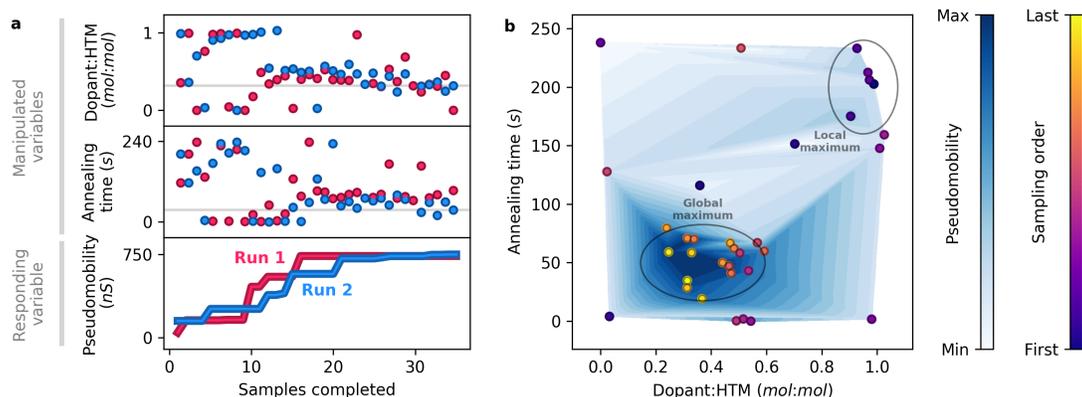


Figure 11.2: Results of thin film pseudomobility optimization carried out by the self-driving lab. (A) Experimental values for cobalt doping ratio, annealing time, and maximum measured pseudomobility as a function of the number of experiments performed for two independent optimization runs. (B) The pseudomobility response surface and sampled points for the second (blue, left) optimization run. The algorithm initially discovered a local maximum, and then discovered the global maximum of the sampled parameter space. Reproduced with permission from Ref. [90].

imum ($\sim 100\%$ doping and annealing time > 200 s) before finding a higher performance region elsewhere in the parameter space. While the eventual rejection of the local maximum confirms that the explore-exploit functionality of PHOENICS can prevent the search from becoming stuck near local optima, we were curious why ADA identified a local maximum at high doping levels. Subsequent investigations of the dark field images of these films revealed annealing-induced dewetting of the films containing moderate amounts of dopant. At elevated doping levels, dewetting was suppressed, allowing a region of improved thermal stability to be identified. The favorable performance of high dopant/high annealing time films was not intuitive,⁶⁶⁴ and was only discovered because the autonomous platform facilitated a search within a broader range of doping and annealing conditions than is typically explored in studies of organic HTMs.

11.3 DISCUSSION

We report here the first use of a self-driving laboratory to optimize composition and processing parameters for thin-film materials. This proof-of-principle study targeted the optimization of a type of thin organic semiconducting film typical to advanced solar cells.

However, the modularity of our robotic platform and control software enables the rapid incorporation of new experiments, techniques, analytical hardware, and algorithms. The ADA platform can, therefore, be tailored for a range of inorganic and organic materials and applications, and even be coupled to automated organic synthesis methodologies developed for the pharmaceutical industry.⁵⁴⁷ We contend that expanding the capabilities of autonomous experimental platforms like ADA will accelerate the optimization of multi-layered materials and devices common to clean energy technologies.

12

Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multi-Component Systems

Apart from minor modifications, this chapter was originally published by Wiley-VCH:

Stefan Langner,* Florian Häse,* José Darío Perea, Tobias Stubhan, Jens Hauch, Loïc M. Roch, Thomas Heumueller, Alán Aspuru-Guzik and Christoph Brabec. Beyond ternary OPV: High-throughput experimentation and self-driving laboratories optimize multi-component systems. *Adv. Mater.* **32**, 1907801 (2020).

Reproduced from Ref. [91] with permission from Wiley-VCH.

ABSTRACT

Fundamental advances to increase the efficiency and stability of organic photovoltaics are achieved by designing ternary blends, which represents a clear trend towards multi-component active layer blends. We report the development of high-throughput and autonomous experimentation methods for the effective optimization of multi-component polymer blends for organic photovoltaics. A method for automated film formation enabling the fabrication of up to 6,048 films per day is introduced. We construct a self-driving laboratory which autonomously evaluates measurements to design and execute the next experiments by equipping this automated experimentation platform with a Bayesian optimization algorithm. To demonstrate the potential of these methods, a four-dimensional parameter space of quaternary organic photovoltaic blends is mapped and optimized for photostability. While with

* These authors contributed equally

conventional approaches roughly 100 mg of material would be necessary, the robot-based platform can screen 2,000 combinations with less than 10 mg and machine learning enabled autonomous experimentation identifies the stable compositions with less than 1 mg.

12.1 INTRODUCTION

With the development of novel non-fullerene acceptors, fundamental performance limitations of fullerene-based organic solar cells (OSCs) have been overcome.⁶⁶⁵ The currently highest performing organic photovoltaics (OPVs) systems are based on PM6:Y6 ternary additives. These additives improve both the charge carrier life time and the mobility of the system to increase efficiencies to 16.5%.^{666,667} Typically, in ternary systems, the ratio of donor to acceptor material (D:A ratio) is kept constant, while only the content of the additive is varied due to limited experimental resources. The existence of other optima beyond these constraints is usually not investigated. However, ternary additives have successfully been introduced to stabilize morphology or prevent oxidation. These desired effects of additives are reverted into detrimental effects depending on the concentration and compatibility of the additive with the host system. An adequate balance is necessary to determine if a given additive can lead to performance enhancements for a given host system. Currently, the Achilles' heel of the highest performing OPV systems is device stability, which motivates the use of a fourth stabilizing additive in performance-optimized ternary systems. Novel experimental methods are required to approach such highly complex optimization tasks with multi-dimensional composition spaces and hundreds of possible candidates for performance-enhancing additives.

These challenges can be addressed with high-throughput experimentation (HTE) to assist the researcher in material synthesis, sample processing, and characterization. Parallelizable high-throughput (HT) methods for polymer samples have been used for adhesion evaluations⁶⁶⁸ and cell biology research.^{669,670} One of the main challenges of combinatorial research in OPV is the scalable fabrication of high-quality individual films. Recent approaches are based on gradient coating or ink-jet printing, which are usually limited by the number of

mixable components, the number of samples or the need for specific ink properties.^{671,672} An HTE coating method for organic semiconductors with a scalable number of mixable components and scalable number of films is not yet established. We introduce a robot-based production of OPV films via drop-casting that can mix a large number of components to form up to 6,048 films of different compositions per day. The reduced material requirements and parallelization capabilities of the HT system allows screening such vast numbers of compositions containing hundreds of possible components at a substantially higher throughput than conventional manual experimentation (see Chapter 9).

Applying this HT system allows screening such a vast number of compositions containing hundreds of possible components that conventional experimental planning and evaluation capabilities are exceeded (see Chapter 9). Therefore, as a second approach, we combine our HTE production line with a machine learning (ML) tool,^{7,48,534,646} to form a self-driving laboratory, for accelerated process optimization and materials discovery. In this ML-driven approach, new experiments are suggested based on all previously collected measurements. To this end, the ML decision-maker infers the outcomes of all possible experiments, leveraging statistical correlations identified from previous measurements, and suggests the most informative ones for future evaluation. The self-driving approach has recently been used to optimize carbon nanotube growth,²⁰³ polyoxometalate clusters, metal-based alloys,⁶¹⁹ and organic synthesis reactions.⁶²⁰ In addition, a thin-film self-driving robot was used to improve film quality and thermal stability of hole-transport materials (HTMs) for clean energy technologies (see Chapter 11).⁹⁰ Further examples are reported and discussed in Chapter 9.

Herein we demonstrate the benefits of HTE and the self-driving approach on the design of photostable material composites for OPVs. While lifetimes of up to 10 years under continuous illumination in nitrogen atmosphere were recently reported,⁶⁷³ the photostability of OSC materials under the influence of oxygen and moisture is a critical challenge. Currently available encapsulation films for OPV materials are a major cost factor for the final product while their water vapor transmission rate is typically $10^{-3} \text{ g m}^{-2} \text{ day}^{-1}$. As higher quality barrier films are currently not manufacturable at a competitive price, optimizing active

layer materials for photostability in the presence of oxygen and water is a critical step for commercial applications. Interestingly, complex photochemical interactions between donor, acceptor, and oxygen can suppress or enhance photooxidation in D:A blends.^{674,675} Adding a third component to the active layer may further increase thermal and light stability.^{676,677} However, the degradation behaviors of higher dimensional multi-component systems may depend on various interactions between the components. Therefore, we use the optimization of four component active layer blends as an experimental proof of concept for our HT film formation and characterization system with ML-enabled self-driving capability. In the following, we compare grid based HTE to the self-driving approach (see Fig. 12.1c).

12.2 HIGH-THROUGHPUT EXPERIMENTATION

The automated platform (see Fig. 12.1b) employed for this study was already used to successfully synthesize and characterize organo metal-halide perovskites and organic nanoparticles with outstanding precision.^{326,678} To form high-quality drop-cast films on inert carriers, we optimized a system to create separate wells on a glass plate. Printing a well structure with UV-curable epoxy allowed a minimal height of the separating walls to reduce material accumulation at the walls resulting in high-quality films with a large homogeneous area (see Fig. 12.1c). To demonstrate the scalability of our approach, we optimized two quaternary blend systems simultaneously. The first system contains PTB7-Th and P3HT as polymer donors (white squares in Fig. 12.1a), while in the second system PBQ-QF is used (black squares in Fig. 12.1a). In both systems, PCBM and oIDTBR are deployed as acceptors. For both material systems, almost 2,100 single films were fabricated and tested within seven days with more than 5,000 absorbance spectra.

In Fig. 12.1c, the experimental workflow of the robotic process is illustrated schematically. Starting with a pre-defined grid of compositions or the closed-loop approach (see Sec. 12.3), individual inks are formulated automatically using a commercially available liquid-handling robot. The inks are drop-cast onto customized 96-well glass substrates to form uniform semiconductor layers. The glass plates are then illuminated with metal halide lamps at

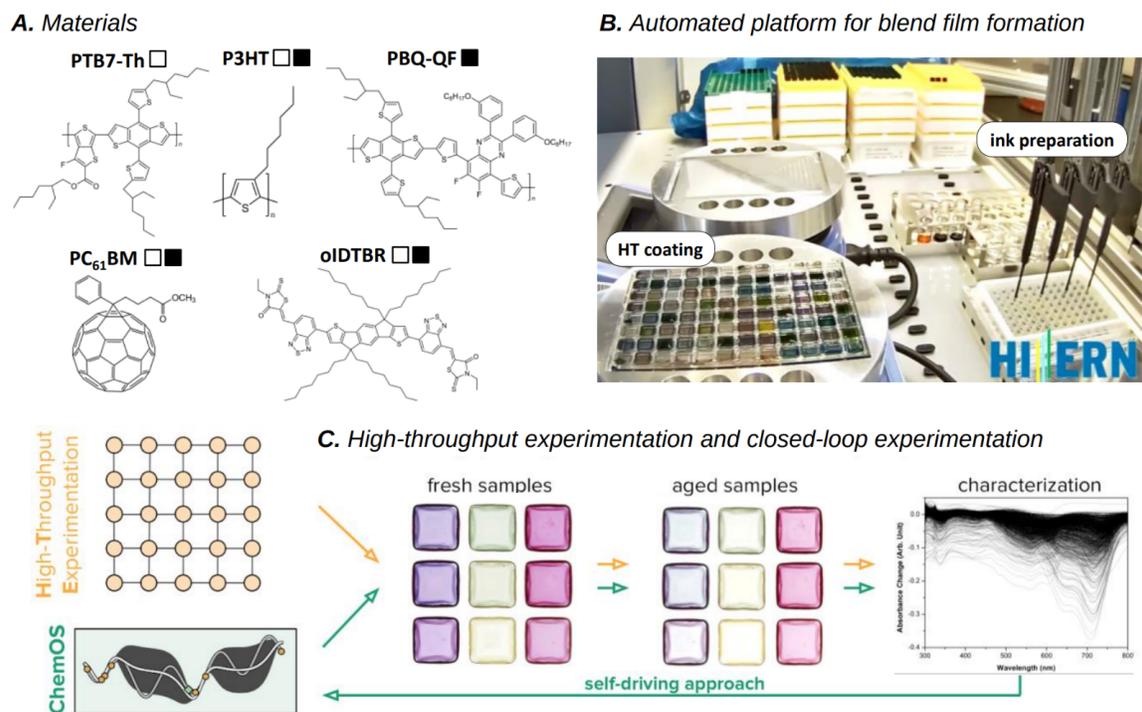


Figure 12.1: (A) Representations of the three polymer donors and the two small molecule acceptors. Note that the first quaternary system consists of P3HT, PBQ-QF, PCBM, and oIDTBr (black square) and the second of P3HT, PTB7-Th, PCBM, and oIDTBr (white square). (B) Side view of the automated platform for ink formulation, coating and characterization. (C) Experimental workflow with the two approaches adopted in this study: conventional high-throughput experimentation via grid and the self-driving approach with the CHEMOS software package. Reproduced from Ref. [91] with permission from Wiley-VCH.

one sun intensity for 18 hours in ambient air to induce photooxidation of the active layers, which can be measured by absorbance loss. Absorbance spectra were recorded before and after light treatment using a microplate reader. The reproducibility of film formation and degradation is very high with an r^2 of 0.97 for 309 tested samples, while the film thickness, as measured by a profilometer, ranged between 70-80 nm. Given the accuracy of our dispensing system, we allowed variations of all four components (PTB7-Th, P3HT, oIDTBR, PCBM) in two percent steps resulting in 23,426 possible compositions. For the first HT test, a subset of 1,041 points was selected containing all single components, binary and ternary variations with 10 wt.-% increment (202 samples) and 820 randomly selected quaternary samples plus 19 duplicates for reproducibility verification. For the second system, the same set of compositions was used while PBQ-QF replaced PTB7-Th. The degradation behavior

of the quaternary system containing PTB7-Th is depicted in Fig. 12.2a (left panel), where degradation is denoted as the integral change of the absorption spectra for each composition before and after degradation. It clearly shows strong degradation of PTB7-Th with a relative absorbance loss of 68 %, while P3HT lost around 19 %. The two acceptor materials exhibited minimal changes in absorbance. Moreover, we observed that the binary series of PTB7-Th toward PCBM or oIDTBR leads to a gradual increase of photostability with acceptor content. But non-linear trends are observed as well, as already slight amounts of P3HT (~ 10 wt.-%) stabilize PTB7-Th completely, leading to blend stabilities similar to pristine P3HT. More surprisingly, blends containing both oIDTBR and PCBM show a drastic destabilization effect causing a large unstable area in the quaternary space (see Fig. 12.2). As this is also observed in a binary mixture of PCBM and oIDTBR (absorbance loss up to 74 %), it seems to be caused by specific interactions between the two acceptors that may be rooted in the formation of particular morphologies that might facilitate the movement of oxygen molecules through the active layer. In the second quaternary system, where PTB7-Th is replaced with PBQ-QF, the same PCBM:oIDTBR instability area is discovered, while most other compositions show improved photostabilities, which is in line with the high stability of pristine PBQ-QF. A detailed viewing of the robot process and the 4D stability-space is found as a cinematic illustration in the appendix of Ref. [91].

Stock solutions of the materials in chlorobenzene were manually prepared with a concentration of 0.6 mg/mL. All further ink formulations were mixed by a liquid handling robot [Freedom Evo 100; Tecan Group AG, Switzerland] using 96-well polypropylene microplates [Eppendorf, Germany]. To guarantee good intermixing, an aspirate/dispense step was repeated for three times before drop-casting 25 μ L of each mixture onto a pre-patterned glass substrate. The films are dried at 60 °C for 4 min. A total of 96 individual films can be formed in 22 min. To obtain stable morphologies, a thermal annealing step at 120 °C for 15 min was performed. Further descriptions are reported in the appendix of Ref. [91].

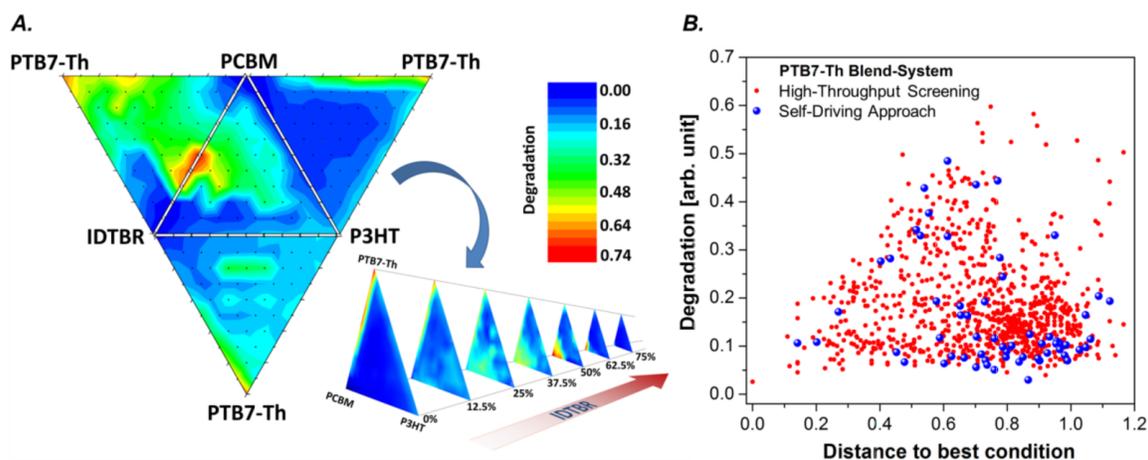


Figure 12.2: (A) Photo-stability of the quaternary system containing PTB7-Th illustrated by the 4D hypersurface and pyramid cross sections. Degradation is defined as the relative spectral loss in absorbance. Black dots represent measured data points, while the color map is interpolated. (B) Comparison the covered experimental space between grid-based high-throughput experimentation and self-driving optimization. The compositional distance, d , to the most stable blend is calculated as $d = \sqrt{x^2 + y^2 + z^2 + t^2}$ with x, y, z, t representing the weight fractions of the four blend components. Reproduced from Ref. [91] with permission from Wiley-VCH.

12.3 AUTONOMOUS EXPERIMENTATION

Instead of a large predefined grid, we data-driven experimental design for the closed-loop approach. The software package CHEMOS^{417,586} coordinates the flow of information between the automated equipment, the researchers, and the experiment planning module (see Chapter 10). The software facilitates the remote exchange of experimental parameters and measurements, which enables the operation of experiment planning strategies and robotics platforms at different locations. In line with the identified steps of closed-loop experimentation,⁷ we extended CHEMOS by an additional module which maps the parameter space of the experiment planner to technical constraints of the robot, which is described in detail in the appendix of Ref. [91]. We used the active learning global optimizer PHOENICS²³³ to learn and navigate the multi-dimensional parameter space (see Chapter 6). This data-driven strategy learns by doing and does not need to be trained with prior measurements. Following a Bayesian optimization strategy, a surrogate model is constructed from parameter kernel densities estimated from a Bayesian neural network (BNN) (see Sec. 5). PHOENICS natively allows for parallel batch-wise experimentation *via* an explicit sampling parameter,

which can suggest experiments explore the candidate space or evaluate the currently most promising candidate. Proposing multiple experiments with different sampling behaviors simultaneously has been shown to accelerate the optimization process and reduce the number of samples needed to identify promising candidates.²³³ In this experiment, CHEMOS leveraged PHOENICS to suggest four blends per closed-loop event, and increases the experimentation throughput by simultaneously coordinating the self-driving approach for the two studied blend systems. To ensure that any active layer composition contains an OPV relevant donor:acceptor ratio, the D:A ratio is limited from 1:5 to 5:1. Starting from 4 random compositions and their stability results, CHEMOS iteratively suggested the next set of 4 compositions that were fabricated, degraded, and characterized. As an initial test run 15 iterations requiring 15 consecutive 18 h degradation tests were made. Here, two identical samples were fabricated for each composition to probe experimental noise. While the self-driving approach focused more on the stable regions in an exploitative search toward a global minimum, an explorative component probed numerous stable and unstable compositions covering largely the same experimental space as the HT test. The experimental space for both HTE and the self-driving approach in the PTB7-Th based system is displayed in Fig. 12.2b. Visualizations of experiments with PBQ-QF are presented in the appendix of Ref. [91]. Here, each data point represents the stability of a quaternary composition plotted over their distance, in terms of compositional difference, to the most stable composition as found by grid-HTE. The results demonstrate how the Bayesian optimization reconstructed the stability distribution of grid-HTE with a much fewer samples (only 7%) compared to HTE. Generally, photostable compounds can be found along the entire distance axis, indicating a broad area of stable compositions rather than a single point minimum. It is observed that the autonomous approach is able to find competitive photostable blends within only 15 learning iterations with 4 samples per iteration. The best compositions found by the data-driven search are similar to the best compositions screened by grid-HTE as shown in the appendix of Ref. [91].

12.4 STATISTICAL COMPARISON AND DISCUSSION

To have a large number of tests that allow statistical comparisons between HTE and the self-driving approach, virtual experiments were performed on a calculated stability grid. From the data collected for the two blend systems during the HTE runs, we constructed probabilistic regression models to emulate the response surfaces of the two experiments. The statistical models, or virtual robots (see Chapter 7), were set up as BNNs and trained to predict the photostability for any possible set of material compositions for the two blend systems. Details on the network architectures, the training procedures and the prediction accuracies on the observed (training sets) and unobserved data (test sets) for both blends are reported in the appendix of Ref. [91].

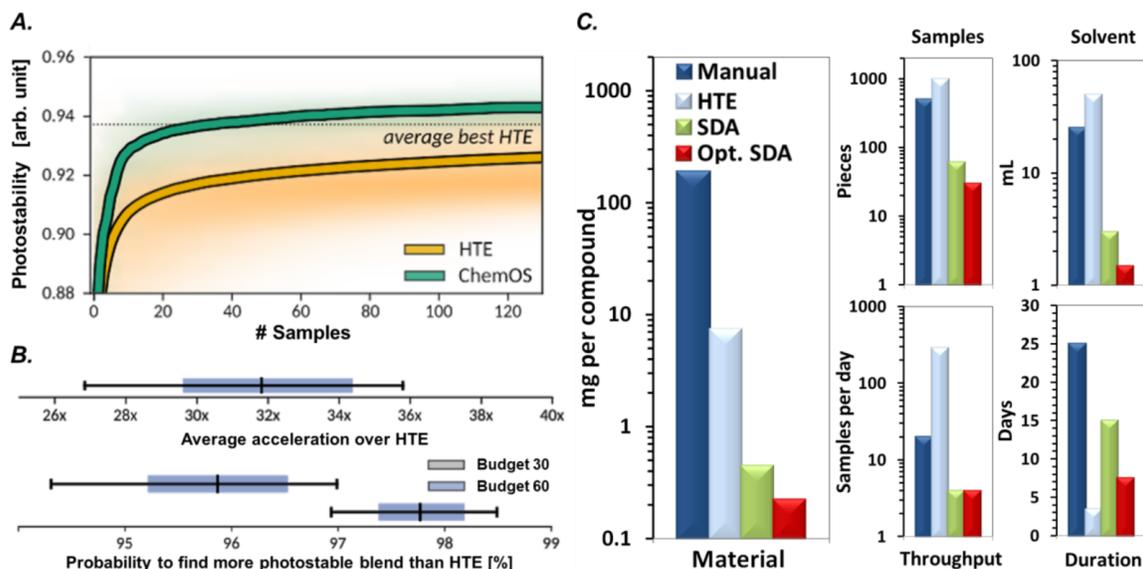


Figure 12.3: (A) Performances of high-throughput experimentation (orange traces) and CHEMOS (green traces) on the virtual robots related to the PTB7-Th based system. Note that the traces were averaged over 10,000 independent grid samplings and 200 independent CHEMOS optimizations. (B) Top: Statistical output of the virtual robot showing the acceleration of autonomous experimentation over HTE; Bottom: Confidence to improve on HTE within given budget of 30 or 60 samples. Both are related to the PTB7-Th based system. (C) Consumption comparison between manual testing, high-throughput experimentation, self-driving approach with budget of 60 samples and virtually optimized self-driving approach with budget of 30 samples. The calculations are limited to one quaternary system. Reproduced from Ref. [91] with permission from Wiley-VCH.

Our results on the virtual robots (Fig. 12.3a) indicate that the autonomous approach requires, on average, 27 samples on the PTB7-Th blend system to identify a blend that is

at least as photostable as the most stable blend discovered by grid-HTE. Furthermore, we find that within a given budget of 30 (60) samples, the autonomous workflow identifies more photostable blends than HTE with a chance of about 96 % (97.5 %) (see Fig. 12.3b). Based on these results, the discovery of photostable blends is accelerated by a factor of ca. 33x for PTB7-Th (and ca. 32x for PBQ-QF, see appendix of Ref. [91]) over conventional HTE as indicated by the virtual robots. Further details on the statistical analysis are reported in the appendix of Ref. [91]. Fig. 12.3c demonstrates the benefits of grid-based HT research and the self-driving approach over manual experimentation. Within this comparison, we assume that systematic manual testing requires the preparation of 500 samples, with a throughput of 20 samples per day, to obtain equivalent information as with HTE of 1,000 samples. Using conventional 1-inch substrates, typical solution concentrations of 30 mg/mL and coating methods, such as spin-coating or blade-coating, a consumption of 188 mg per compound is expected. In contrast, HTE needs only 7.5 mg per compound due to the low concentration of 0.6 mg/mL used in drop-casting. Moreover, even as 288 samples are tested in a single day, HTE guarantees a stable process with highly reproducible data. The substantially reduced number of samples for the self-driving approach with only 60 samples further reduces the amount of consumed materials to 0.45 mg per compound. Following the findings from the virtual robots, the autonomous approach identifies the most photostable blends consistently after about 30 candidate evaluations (0.225 mg). Note that for CHEMOS the given throughput of 4 samples per day and blend system is based on the iterative degradation process. Nevertheless, by parallelization across blend systems, CHEMOS is able to run and optimize 16 blend systems simultaneously, further accelerating the discovery process.

12.4.1 VIRTUAL ROBOT

The statistical significance of the experimentally obtained performance difference between the autonomous approach and HTE was assessed in detail with the construction of virtual robots. Virtual robots can be designed *via* probabilistic ML models to emulate an experi-

mental procedure *in silico*. Assuming that the virtual robot reproduces the measurements of previous experiments sufficiently well, the merit of new experiments can be estimated computationally without conducting further experiments. Consequently, virtual robots can serve as a benchmark to determine performance differences between different experiment planning strategies. We construct virtual robots from BNNs, which are trained to reproduce the photodegradation measured for individual polymer blends in the HTE approach. BNNs constitute probabilistic ML models which can be trained *via* variational expectation-maximization in a Bayesian framework. Thus, BNNs are intrinsically robust to overfitting and can infer both the expected photodegradation and the degree of experimental noise to resemble experimental conditions. A total of 850 experiments (81.7%) were randomly selected from the 1,041 HTE experiments for both blend systems to train BNNs with 5-fold cross-validation. The remaining 190 experiments (18.3%) were used as a test set to determine the generalization performance of the trained BNNs. BNNs were constructed as fully connected feedforward networks with three hidden layers. To account for the fact that the photodegradation of any given polymer blend cannot be negative, we selected ReLU activation functions for the output layer. Weights and biases were modeled with Gaussian priors. Additional network hyperparameters and the selected values are reported in Tab. 12.1.

Table 12.1: Hyperparameters for the Bayesian neural networks used as virtual robots to emulate the photodegradation of individual polymer blends.

Hyperparameter	Selected value
Number of hidden layers	3
Neurons per hidden layer	120
Hidden layer activations	Leaky ReLU
Learning rate	0.001
Dropout rate	0.2

Both polymer blend ratios and associated photo degradations were rescaled prior to being presented to the BNNs. Polymer blend ratios were transformed from the 4-simplex to the three-dimensional unit cube, while individual photodegradations were divided by the average photodegradation taken over the 850 experiments used for cross-validation. Prediction accuracies for the BNNs trained on the two blend systems are illustrated in Fig. 12.4. We

observe that the BNNs are capable of reproducing the experimental measurements of the test sets with coefficients of determination of 0.88 (0.87) on the PBQ-QF (PTB7-Th) blend system. The agreement of coefficients of determination between the test and the training sets indicates good transferability of BNNs.

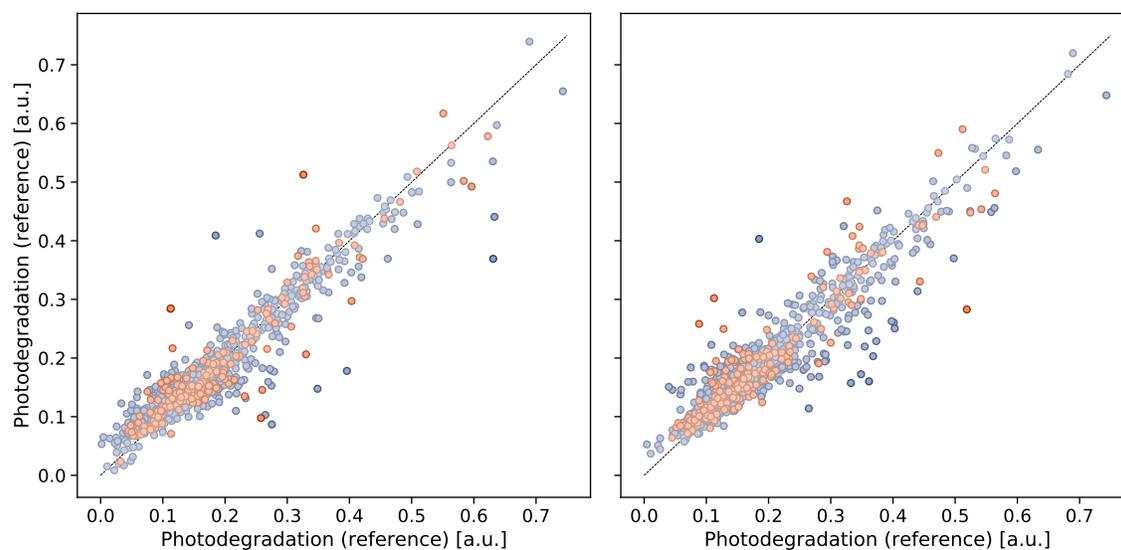


Figure 12.4: Prediction accuracies of the virtual robots constructed from Bayesian neural networks on the two studied blend systems. Predicted photo-degradations are compared to measurements obtained from high-throughput experimentation. Reproduced from Ref. [91] with permission from Wiley-VCH.

12.5 CONCLUSIONS AND OUTLOOK

We have demonstrated how complex, multi-component optimizations of next-generation OSC active layers can be substantially accelerated by novel HT film deposition techniques combined with automated characterization and ML-driven experimental design. Over 2000 different quaternary active layers were tested in seven days with a materials consumption of less than 15 mg per component. This throughput allows a more thorough stability investigation of novel OPV materials and their interactions with other blend components at ultra-low material consumptions. Implementing a ML optimization algorithm, equivalent information about stability minima and maxima could be found with a sample reduction of around 93%. In our specific test case, P3HT and PBQ-QF rich blends show improved

stability over PTB7-Th rich blends. PCBM and oIDTBR can destabilize each other with dramatic consequences for the active layer blend. As a next step, the inclusion of additional characterization methods and target properties, *e.g.*, photoluminescence, or conductivity, could allow to not only optimize for stability but also electrical performance.

Conclusions and future directions

This page is intentionally left blank.

13

Summary and outlook

Data-driven technologies are emerging as versatile and viable tools for scientific research. While statistical modeling has always been essential to scientific investigations,^{45–47} the black-box applicability of machine learning (ML) models has recently experienced increased interest across several fields for tasks where the explicit structure of the data-generating system is unknown.^{48–60} With the capacity to identify patterns from data and to leverage statistical correlations in the absence of explicit physical assumptions, ML methods have the potential to connect thorough bottom-up theoretical models with Edisonian trial-and-error strategies in scientific discovery workflows. Especially for the discovery of functional molecules and advanced materials for clean energy technologies, where comprehensive theoretical models and experimental approaches are often time-consuming and resource-demanding, data-driven strategies can enable far-reaching insights to discover solutions to some of the immediate societal challenges of this century. Succeeding in this endeavor can promote our understanding of some of the fundamental open challenges around artificial light harvesting. Revealing the interplay between microscopic features and mesoscale properties of natural pigment-protein complexes, the quantum mechanical effects modulating non-radiative decay rates in organic solar cells (OSCs), or the influence of processing conditions on the stabilities and efficiencies of perovskite materials could catalyze the transition to a low-carbon economy. Already to date, clean energy research increasingly benefits from data-driven tools at all stages of scientific investigation, including property screening of functional molecules,^{66,67} materials candidate selection for light-harvesting devices,^{68–70,510} analysis of empirical measurements,⁷¹ and the identification and interpretation of phenomenological

trends describing the emergence of device properties from materials compositions and molecular structures.⁷³ By highlighting causal structure-property relations from experiments and empirical evidence, data-driven approaches can extend the current boundaries for cutting edge experimental and computational technologies for both device conception and device characterization. In this spirit, we have explored the benefits and advantages of data-driven strategies to enabling new routes to the discovery of clean energy technologies,^{81,90,417,586?} ranging from the conception of hypotheses³³⁰ over the design of experiments^{233–235} and the analysis of collected measurements³⁸⁵ to the interpretation of experimental outcomes.⁷² Designing useful data-driven tools for scientific investigation requires a thorough understanding of the studied physical processes. In contrast to the more traditional fields of ML research, such as natural language processing (NLP) or image recognition, scientific applications pose unique challenges to data-driven tools. Scientific discovery tasks typically involve relatively small datasets that contain a noticeable degree of inherent noise. The studied physical systems might also obey particular symmetries, and all structure-property relations for molecular systems are expected to be governed by the known laws of physics. Finally, science strives to conceptualize the governing principles of nature. To advance scientific research for artificial light-harvesting and beyond, the requirements on ML models for scientific discovery go beyond high prediction accuracies and necessarily need to extend to balancing flexibility, robustness, and interpretability. The transition to predictive, intuitive, and interpretable ML models to complement theoretical studies and experimental investigations has the potential to provide crucial scientific insights and inspire design rules to improve materials, processing conditions, and eventually devices for clean energy technologies.

13.1 CONCLUSIONS

In this dissertation, we have explored how data-driven approaches can enable new research strategies within the framework of the scientific method. With the ever-growing need to transition to a low-carbon economy, we primarily focused our efforts on data-driven tools for the discovery of light-harvesting materials and the understanding of light-matter interactions.

Part I discussed the integration of data-driven methods into existing workflows for scientific investigation. We demonstrated how ML models can help to answer questions that were too resource-demanding to be addressed with state-of-the-art experimental and computational techniques. We further evaluated how ML models allow to rephrase these questions and even enable new routes to scientific investigations for clean energy technologies. Chapter 3 introduced fully connected feed-forward neural networks for the prediction of excitation energy transfer (EET) properties of excitonic systems as they are frequently encountered in natural photosynthesis, dye-sensitized solar cells, and organic photovoltaics (OPVs). We found that data-driven approaches can predict EET properties at a high degree of accuracy compared to Markovian approximations to the numerically exact physical model, but at a much more favorable computational demand. The ML models constructed in this chapter show how large-scale studies of the structure-property relations in excitonic networks can be enabled to inspire design choices for novel excitonic devices. Chapter 4 presented an approach to study highly conductive polymers, which are used as transparent contacts in OPV devices such as bulk-heterojunction solar cells. Understanding the relation between the molecular organization of these materials and the emergence of their conductive properties from their microscopic structures is a critical challenge for practical applications. Yet, direct experimental measurements of the conductivity are inherently demanding. We demonstrated that Bayesian convolutional neural networks (CNNs) can be used to estimate the electronic coupling of organic electronics materials purely based on their absorption spectra. In this spirit, this chapter highlighted the possibility to infer statistical correlations between electronic and optical properties to enable less challenging experimental investigations. Our findings suggest that the use of proxy measurements in combination with data-driven tools has the potential to facilitate a whole new set of experimental approaches to characterize advanced materials for artificial light-harvesting. Finally, Chapter 5 outlined how ML can spark physical insights and scientific understanding. We considered the chemiluminescent dissociation of 1,2-dioxetane and illustrated how data-driven strategies could evidence the prevalent nuclear motions which promote or delay dissociation. Our analyses revealed the

dissociation mechanism from a sparse set of simulated dissociation trajectories. This understanding can inspire further design choices to manipulate the chemiluminescence yield of dioxetane.

In Part II, we focused on a narrower context of scientific discovery and introduced a set of algorithmic tools for data-driven experiment planning in closed-loop processes. This approach has been identified to be of particular relevance for tasks in physics, chemistry, and materials science, where experiments can be time-consuming and resource-demanding. The closed-loop approach aims to identify specific values for controllable parameters that yield the desired experimental outcomes. Controllable parameters typically encountered in light-harvesting research include continuous variables such as temperatures or amounts of solvents and materials, but could also involve categorical design choices, including the choice of a particular material or constituent. In addition, viable device solutions are often optimized for not just one objective but multiple objectives at once. In Chapter 6, we introduced a Bayesian optimization strategy capable of identifying optimal process parameters in closed-loop processes. This algorithm, which we called PHOENICS, natively supports batch-wise optimizations and requires fewer parameter evaluations than previously deployed strategies at a more favorable computational scaling. Chapter 7 introduced GRYFFIN, which extends this optimization approach to categorical variables. While being a competitive categorical optimizer compared to established approaches, GRYFFIN can also accelerate the search by leveraging domain knowledge in the form of physicochemical descriptors. It can further refine these descriptors to inspire design choices and promote scientific understanding. Finally, we developed CHIMERA in Chapter 8 to enable flexible multi-objective optimization in closed-loop processes requiring limited *a priori* preference information while keeping the number of experiments low. All three of these algorithms have been designed to serve as algorithmic experiment planning strategies to fuel autonomous experimentation workflows for scientific discovery. The extensive benchmarks presented in this part suggest that the introduced algorithms can identify viable materials candidates and processing conditions with only limited feedback, thus enabling the cost-efficient exploration of large design spaces. We

have also demonstrated that GRYFFIN and CHIMERA can inspire design choices and promote scientific understanding while navigating the design space.

Finally, in Part III, we illustrated the benefits of autonomous experimentation to clean energy research on two prototypic studies targeted the discovery of conductive thin-films for perovskite solar cells (PSCs) and the discovery of photostable OPVs. At the beginning of this part, we discussed the potential advantages and challenges of autonomous platforms as a next-generation approach to experimentation in Chapter 9. Based on these considerations, we introduced a comprehensive software package, CHEMOS, in Chapter 10. CHEMOS was designed to simplify the implementation and deployment of autonomous platforms by enabling the integration of algorithmic strategies for data-driven experiment planning with robotic platforms for automated experimentation. In Chapter 11, we presented a prototype of an autonomous experimentation platform for the discovery of thin-film materials, which uses the algorithmic tools introduced in Chapters 6 and 10. We found that, in addition to enabling experimentation without human intervention, our prototype platform evidenced unexpected experimental outcomes regarding the conductivity of thin-film materials. This observation suggests that autonomous experimentation has the capacity to further inspire scientific insights. Chapter 12 presented another autonomous platform based on the same algorithmic tools which targets the discovery of photostable polymer blends for photovoltaic applications. OPV are typically composed of blends of organic donor and acceptor materials, and multi-component active layer blends have recently been suggested to increase both the efficiency and stability of OPVs. However, the vast number of possible materials combination poses a challenge to the discovery of viable multi-component materials. We have shown that these large, intractable design spaces can be efficiently navigated with autonomous platforms, which significantly reduce the resource demands of systematic investigations. As such, we suggest that autonomous platforms can enable large-scale research on more complex, and thus more advanced, clean energy technologies.

As part of this dissertation, several algorithms have been conceived, developed, and made available under permissive licenses in public repositories on GitHub: PHOENICS,⁷⁸

CHIMERA,⁸⁰ GRYFFIN,⁷⁹ and CHEMOS.⁶⁴⁴ Given the range of applications on which these algorithms have been demonstrated to date and inspired by the initial successes demonstrated in Chapters 11 and 12 we hope that these algorithmic tools can contribute to the implementation of next-generation approaches to experimentation with autonomous platforms. We believe that data-driven strategies have the potential to advance scientific research beyond conventional approaches and invite the scientific community to join us in this venture and push the frontiers of clean energy research.

13.2 FUTURE DIRECTIONS

By the time this dissertation is written, the data-driven strategies for experiment planning developed in Chapters 6, 7, 8, and the tools for autonomous research introduced in Chapter 10 are further extended and used to enable several discovery projects for diverse applications including clean energy technologies and others.

Following the encouraging first results on the autonomous optimization of Suzuki coupling reactions discussed in Chapter 7, the group of Prof. Jason Hein at the University of British Columbia, Vancouver, develops an end-to-end solution for autonomous experimentation. This integrated platform targets the streamlined identification of promising ligands and processing conditions for Suzuki reactions of pharmaceutical interest in a closed-loop process with real-time synthesis and characterization of the reaction products. Contrary to the prototype study highlighted in Chapter 7, this collaboration focuses on the maximization of the specificity of the reaction. With this goal in mind, the autonomous workflow is designed to promote the formation of the desired product while suppressing the formation of other, undesired products. In the autonomous workflow, we access a much broader set of ligands to thoroughly analyze the benefits of individual ligands to the selectivity of the reaction. Initial results suggest that previously unidentified ligands yield high reaction selectivities. This observation highlights the capacity of autonomous systems to evidence unexpected outcomes, similar to the observations in Chapter 11. We believe that the autonomous workflow designed in this study will inspire further design choices for highly selective ligands

for medicinal chemistry.

Another extension of this work, in collaboration with the group of Prof. Daniel Tabor at the Texas A&M University, targets the integration of GRYFFIN with high-throughput (HT) virtual screening techniques for the discovery of clean energy storage solutions. This collaboration focuses on flow battery electrolyte discovery. The molecules for these batteries must satisfy several objectives, *e.g.*, correct cell potential, long term stability, high solubility, and synthesizability, and the space of possible molecules is enormous. Our ongoing work seeks to accelerate the brute force search for promising flow battery electrolytes by building these molecules out of modular functional group building blocks. Our work focuses on finding the best representation for these underlying building blocks, following both traditional cheminformatics fingerprint techniques and physically-motivated electronic descriptors. Based on these initial considerations, we further examine the results of several parallel runs of the optimization procedure in this large chemical space to gain further physical insights into how to construct the next candidate space. This project is still underway, but to date, has served as a prototypical example of how a domain scientist’s expertise can be further leveraged both on the front end and the back end of data-driven applications to chemical and materials discovery.

Finally, the algorithmic tools developed in this dissertation are being integrated into *in silico* autonomous workflows for the discovery of organic semiconducting lasers. Organic lasers are broadly tunable coherent sources which can be manufactured at low cost and are well suited for lab-on-a-chip applications. These π -conjugated systems are expected to require very low pumping energies, which make them promising candidates to enable spectroscopy, chemical sensing, and telecommunication. Similarly to the discovery of flow battery electrolytes, the space of possible laser candidates is vast and intractable with conventional strategies to materials discovery. Our workflow targets the autonomous identification of potential laser candidates based on their optical properties, conductivity, solubility, and stability.

These three efforts present only a few examples of the transition towards autonomous dis-

covery which the chemistry and materials science communities join more and more. The urge to streamline the discovery process is imminent, particularly for clean energy technologies, and data-driven tools seem to alleviate some of the limitations of established approaches. Particularly in the absence of tractable physical models and inexpensive experimentation protocols as commonly encountered, *e.g.*, in clean energy research, the statistical nature of ML models has shown promise to amplify cutting edge theoretical tools and experimental technologies to enable more comprehensive discovery workflows. In this spirit, we have demonstrated that the algorithms introduced in Chapters 6, 7, 8, and 10 contribute to the necessary steps towards the advancement of science with next-generation autonomous platforms.

Although the successes of autonomous experimentation platforms have recently been reported and promising data-driven tools to reach autonomy are discussed, current prototypes present at least one substantial drawback: the inability to dynamically derive and develop transferable domain expertise from collected empirical evidence. Current data-driven experiment planning strategies, including those presented in this dissertation, start the experimentation process without scientifically motivated assumptions about the causal relation between experimental parameters and responses. However, precisely supplied prior information or physically motivated expectations can guide an experiment planning algorithm towards prioritizing the execution of likely promising experiments. Prior knowledge could reduce information redundancies in the proposed experiments and thus streamline the scientific discovery process. Human experimenters implicitly leverage such prior knowledge in the form of intuition or domain expertise, which they gradually accumulate over many years of practicing science. Following their intuition, humans can be remarkably accurate on a qualitative level when estimating the properties of molecules, inferring the process parameters for an experiment, or even choosing promising hyperparameters for a deep learning model. Intuitive artificial systems that can formalize and exploit general trends observed across several applications have the potential to infer materials properties with fewer data. The scientific conceptualization of this accumulated data-driven domain expertise can po-

tentially inspire physical insights. With this opportunity, machine-driven intuition directly contributes to the design of interpretable ML models and fosters scientific understanding.

Finally, significant benefits to scientific research can only be achieved if the advantages and strengths of data-driven approaches can be translated successfully into existing laboratory environments at low deployment costs and obstacles. By leveraging empirical correlations, data-driven models intend to be useful for a given task. To that end, it is not sufficient to stop at the construction of a highly predictive and interpretable data-driven tool. The utility of such a model further needs to be promoted by making it accessible. Ultimately, the degree of success or failure of data-driven approaches in the physical and life sciences will be determined by how well they support researchers in their daily routines and how well they integrate into common laboratory environments. The work in this dissertation presents steps towards the advancement of clean energy research with synergistic integrations of data-driven approaches into discovery workflows. The campaign to recognize and leverage the benefits data-driven approaches to their fullest is only at its beginning.

Bibliography

- [1] E.L. Opie. Diabetes mellitus associated with hyaline degeneration of the islands of Langerhans of the pancreas. *Johns Hopkins Hospital Bulletin*, (125), 1901.
- [2] E.L. Opie. On the relation of chronic interstitial pancreatitis to the islands of Langerhans and to diabetes mellitus. *J. Exp. Med.*, 5(4):397–428, 1901.
- [3] F.G. Banting and C.H. Best. Pancreatic extracts. *J. Lab. Clin. Med.*, 7(8):464–472, 1922.
- [4] M. Bliss. Rewriting medical history: Charles Best and the Banting and Best myth. *J. Hist. Med. All. Sci.*, 48(3):253–274, 1993.
- [5] J. Bardeen and W.H. Brattain. The transistor, a semi-conductor triode. *Phys. Rev.*, 74(2):230, 1948.
- [6] J. Deisenhofer, O. Epp, K. Miki, R. Huber, and H. Michel. Structure of the protein subunits in the photosynthetic reaction centre of rhodospseudomonas viridis at 3Å resolution. *Nature*, 318(6047):618–624, 1985.
- [7] D.P. Tabor, L.M. Roch, S.K. Saikin, C. Kreisbeck, D. Sheberla, J.H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C.J. Brabec, B. Maruyama, K.A. Perrson, and A. Aspuru-Guzik. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.*, 3:5–20, 2018.
- [8] B. Dunn, H. Kamath, and J.M. Tarascon. Electrical energy storage for the grid: a battery of choices. *Science*, 334(6058):928–935, 2011.
- [9] X. She, A.Q. Huang, and R. Burgos. Review of solid-state transformer technologies and their application in power distribution systems. *IEEE Trans. Emerg. Sel. Topics Power Electron.*, 1(3):186–198, 2013.
- [10] T.M.I. Mahlia, T.J. Saktisahdan, A. Jannifar, M.H. Hasan, and H.S.C. Matseelar. A review of available methods and development on energy storage; technology update. *Renew. Sust. Energy. Rev.*, 33:532–545, 2014.
- [11] V. Chabot, D. Higgins, A. Yu, X. Xiao, Z. Chen, and J. Zhang. A review of graphene and graphene oxide sponge: material synthesis and applications to energy and the environment. *Energy Environ. Sci.*, 7(5):1564–1596, 2014.
- [12] A. Ferreira, B.L. Duarte, P.R.O. Nova, and A.R. Marques. Multifunctional material systems: a state-of-the-art review. *Comp. Struct.*, 151:3–35, 2016.
- [13] J.R. Werber, C.O. Osuji, and M. Elimelech. Materials for next-generation desalination and water purification membranes. *Nat. Rev. Mater.*, 1(5):16018, 2016.

-
- [14] S. Priya, H.C. Song, Y. Zhou, R. Varghese, A. Chopra, S.G. Kim, I. Kanno, L. Wu, D.S. Ha, J. Ryu, et al. A review on piezoelectric energy harvesting: Materials, methods, and circuits. *Energy Harvesting and Systems*, 4(1):3–39, 2019.
- [15] J.L. Brédas, E.H. Sargent, and G.D. Scholes. Photovoltaic concepts inspired by coherence effects in photosynthetic systems. *Nat. Mater.*, 16(1):35, 2017.
- [16] T. Chatterjee and K.T. Wong. Perspective on host materials for thermally activated delayed fluorescence organic light emitting diodes. *Adv. Opt. Mater.*, 7(1):1800565, 2019.
- [17] H. Andersen and B. Hepburn. Scientific method. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2016 edition, 2016.
- [18] W.K. Röntgen. Über eine neue Art von Strahlen: Vorläufige Mitteilung. *Sitzungsber. Phys. Med. Gesell.*, 1895.
- [19] A. Fleming. Penicillin. Its practical application. *Penicillin. Its practical application.*, 1946.
- [20] P. Godfrey-Smith. *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press, 2009.
- [21] T.A. Brody. *The philosophy behind physics*. Springer, 1993.
- [22] T.S. Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.
- [23] P. Galison. *How experiments end*. University of Chicago Press, 1987.
- [24] P. Harrison, R.L. Numbers, and M.H. Shank. *Wrestling with nature: from omens to science*. University of Chicago Press, 2011.
- [25] R. Hilliam. *Galileo Galilei: Father of modern science*. The Rosen Publishing Group, Inc, 2004.
- [26] S. Drake. *Galileo at work: his scientific biography*. Courier Corporation, 2003.
- [27] I. Newton. Philosophiæ naturalis principia mathematica (mathematical principles of natural philosophy). *London (1687)*, 1687, 1987.
- [28] W. Wien. Untersuchungen über die elektrische Entladung in verdünnten Gasen. *Ann. Phys.*, 301(6):440–452, 1898.
- [29] W. Kaufmann. Die magnetische und elektrische Ablenkbarkeit der Bequerelstrahlen und die scheinbare Masse der Elektronen. *Nach. v. d. k. Gesellsch. d. Wissensch. Math.-phys. Kl. Götting.*, 1901:143–155, 1901.
- [30] G.F.C. Searle. On the steady motion of an electrified ellipsoid. *Proc. Phys. Soc.*, 15(1):264, 1896.

- [31] H.A. Lorentz. Über die scheinbare Masse der Ionen. *Phys. Z.*, 2:78–80, 1901.
- [32] M. Abraham. Dynamik des Elektrons. *Nach. v. d. k. Gesellsch. d. Wissensch. Math.-phys. Kl. Götting.*, 1902:20–41, 1902.
- [33] H.A. Lorentz. Electromagnetic phenomena in a system moving with any velocity smaller than that of light. In *Royal Netherlands Academy of Arts and Sciences, Proceedings*, pages 809–831. 1904.
- [34] H. Poincaré. Sur la dynamique de l'électron. *Rend. Circ. Mat. Palermo*, 21:129–176, 1906.
- [35] W. Kaufmann. Über die elektromagnetische Masse des Elektrons. *Nach. v. d. k. Gesellsch. d. Wissensch. Math.-phys. Kl. Götting.*, 1902:291–296, 1902.
- [36] M. Planck. Die Kaufmannschen Messungen der Ablenkbarkeit der β -Strahlen in ihrer Bedeutung für die Dynamik der Elektronen. *Phys. Z.*, 1906.
- [37] A. Einstein. Zur Elektrodynamik bewegter Körper. *Ann. Phys.*, 322(10):891–921, 1905.
- [38] A.H. Bucherer. Die experimentelle Bestätigung des Relativitätsprinzips. *Ann. Phys.*, 333(3):513–536, 1909.
- [39] G. Neumann. Die träge Masse schnell bewegter Elektronen. *Ann. Phys.*, 350(20):529–579, 1914.
- [40] M.Y. Vardi. Science has only two legs. *Comm. ACM*, 53(9):5–5, 2010.
- [41] K. Lindorff-Larsen, S. Piana, R.O. Dror, and D.E. Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.
- [42] R. Highfield. Large hadron collider: Thirteen ways to change the world. *The Daily Telegraph. London. Retrieved*, pages 10–10, 2008.
- [43] S. Wozny, M. Yang, A.M. Nardes, C.C. Mercado, S. Ferrere, M.O. Reese, W. Zhou, and K. Zhu. Controlled humidity study on the formation of higher efficiency formamidineum lead triiodide-based solar cells. *Chem. Mater.*, 27(13):4814–4820, 2015.
- [44] E. Bracken. Combating humidity-the hidden enemy in manufacturing. *Sens. Rev.*, 1997.
- [45] H.B. Prosper. Practical statistics for particle physicists. *arXiv preprint arXiv:1608.03201*, 2016.
- [46] K. Cranmer. Practical Statistics for the LHC. *arXiv preprint arXiv:1503.07622*, 2015.
- [47] G. Cowan. Statistics for Searches at the LHC. In *LHC Phenomenology*, pages 321–355. Springer, 2015.
- [48] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, and A. Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547, 2018.

-
- [49] Ankit Agrawal and Alok Choudhary. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.*, 4(5):053208, 2016.
- [50] K. Rajan. Materials informatics: The materials “gene” and big data. *Annu. Rev. Mat. Res.*, 45:153–169, 2015.
- [51] Y. Liu, T. Zhao, W. Ju, and S. Shi. Materials discovery and design using machine learning. *J. Materiomics*, 3(3):159–177, 2017.
- [52] L. Zdeborová. New tool in the box. *Nat. Phys.*, 13:420–421, 2017.
- [53] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim. Machine learning in materials informatics: Recent applications and prospects. *npj Comput. Mater.*, 3(1):54, 2017.
- [54] L. Ward and C. Wolverton. Atomistic calculations and materials informatics: A review. *Curr. Opin. Solid State Mater. Sci.*, 21(3):167–176, 2017.
- [55] T. Mueller, A.G. Kusne, and R. Ramprasad. Machine learning in materials science: Recent progress and emerging applications. *Rev. Comp. Chem.*, 29:186–273, 2016.
- [56] A. Zunger. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.*, 2(4):0121, 2018.
- [57] J.E. Gubernatis and T. Lookman. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Phys. Rev. Mater.*, 2(12):120301, 2018.
- [58] B.R. Goldsmith, J. Esterhuizen, J.X. Liu, C.J. Bartel, and C.A. Sutton. Machine learning for heterogeneous catalyst design and discovery. *AIChE-Journal*, 64(7):2311–2323, 2018.
- [59] K. Takahashi and Y. Tanaka. Materials informatics: A journey towards material design and synthesis. *Dalton Trans.*, 45(26):10497–10499, 2016.
- [60] A.P. Bartók, R. Kondor, and G. Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):184115, 2013.
- [61] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural. Inf. Process. Syst.*, pages 1097–1105, 2012.
- [62] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [63] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

- [64] D. Bzdok, N. Altman, and M. Krzywinski. Statistics versus machine learning. *Nat. Methods*, 15(4):233–234, 2018.
- [65] F. Li, X. Peng, Z. Wang, Y. Zhou, Y. Wu, M. Jiang, and M. Xu. Machine learning (ML)-assisted design and fabrication for solar cells. *Energy Environ. Mater.*, 2(4):280–291, 2019.
- [66] L. Weston and C. Stampfl. Machine learning the band gap properties of kesterite $I_2-II-IV-V_4$ quaternary compounds for photovoltaics applications. *Phys. Rev. Mater.*, 2(8):085407, 2018.
- [67] K. Choudhary, M. Bercx, J. Jiang, R. Pachter, D. Lamoen, and F. Tavazza. Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods. *Chem. Mater.*, 31(15):5900–5908, 2019.
- [68] X. Yang, J. Zhang, K. Yoshizoe, K. Terayama, and K. Tsuda. Chemts: An efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.*, 18:972–976, 2017.
- [69] M. Sumita, X. Yang, S. Ishihara, R. Tamura, and K. Tsuda. Hunting for organic molecules with artificial intelligence: Molecules optimized for desired excitation energies. *ACS Cent. Sci.*, 4(9):1126–1133, 2018.
- [70] P.B. Jørgensen, M. Mesta, S. Shil, J.M. García Lastra, K.W. Jacobsen, K.S. Thygesen, and M.N. Schmidt. Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.*, 148(24):241735, 2018.
- [71] T. Chen, Y. Zhou, and M. Rafailovich. Application of machine learning in perovskite solar cell crystal size distribution analysis. *MRS Adv.*, 4(14):793–800, 2019.
- [72] F. Häse, I.F. Galván, A. Aspuru-Guzik, R. Lindh, and M. Vacher. How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry. *Chem. Sci.*, 10(8):2298–2307, 2019.
- [73] J. Im, S. Lee, T.W. Ko, H.W. Kim, Y.K. Hyon, and H. Chang. Identifying Pb-free perovskites for solar cells by machine learning. *npj Comput. Mater.*, 5(1):1–8, 2019.
- [74] B. Sanchez-Lengeling and A. Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.
- [75] Z.M. Sherman, M.P. Howard, B.A. Lindquist, R.B. Jadrich, and T.M. Truskett. Inverse methods for design of soft materials. *J. Chem. Phys.*, 152(14):140902, 2020.
- [76] J. Noh, J. Kim, H.S. Stein, B. Sanchez-Lengeling, J.M. Gregoire, A. Aspuru-Guzik, and Y. Jung. Inverse design of solid-state materials via a continuous representation. *Matter*, 1(5):1370–1384, 2019.
- [77] J.G. Freeze, H.R. Kelly, and V.S. Batista. Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. *Chem. Rev.*, 119(11):6595–6612, 2019.

-
- [78] F. Häse. Phoenix: A Bayesian optimizer for chemistry. <https://github.com/aspuru-guzik-group/phoenix>, 2018.
- [79] F. Häse. Gryffin: An algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications to chemistry. <https://github.com/aspuru-guzik-group/gryffin>, 2018.
- [80] F. Häse. Chimera: enabling hierarchy-based multi-objective optimization for self-driving laboratories. <https://github.com/aspuru-guzik-group/phoenix>, 2018.
- [81] F. Häse, L.M. Roch, and A. Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. *Trends Chem.*, 1(3):282–291, 2019.
- [82] K.F. Jensen, C.W. Coley, and N.S. Eyke. Autonomous discovery in the chemical sciences part I: Progress. *Angew. Chem. Int. Ed.*, 2019.
- [83] H.S. Stein and J.M. Gregoire. Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. Sci.*, 10(42):9640–9649, 2019.
- [84] C.W. Coley, N.S. Eyke, and K.F. Jensen. Autonomous discovery in the chemical sciences part II: Outlook. *Angew. Chem. Int. Ed.*, 2019.
- [85] R.K. Pachauri, M.R. Allen, V.R. Barros, J. Broome, W. Cramer, R. Christ, J.A. Church, L. Clarke, Q. Dahe, P. Dasgupta, et al. *Climate change 2014: Synthesis report. Contribution of working groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Ipc, 2014.
- [86] J.M. Koh. Green infrastructure financing. *Springer Books*, 2018.
- [87] D.G. Nocera. The artificial leaf. *Acc. Chem. Res.*, 45(5):767–776, 2012.
- [88] N.S. Lewis and D.G. Nocera. Powering the planet: Chemical challenges in solar energy utilization. *Proc. Natl. Acad. Sci.*, 103(43):15729–15735, 2006.
- [89] E. Maine and E. Garnsey. Commercializing generic technology: The case of advanced materials ventures. *Res. Policy*, 35(3):375–393, 2006.
- [90] B.P. MacLeod, F.G.L. Parlane, T.D. Morrissey, F. Häse, L.M. Roch, K.E. Dettelbach, R. Moreira, L.P.E. Yunker, M.B. Rooney, J.R. Deeth, V. Lai, G.J. Ng, H. Situ, R.H. Zhang, M.S. Elliott, T.H. Haley, D.J. Dvorak, A. Aspuru-Guzik, J.E. Hein, and C.P. Berlinguette. Self-driving laboratory for accelerated discovery of thin-film materials. *arXiv preprint arXiv:1906.05398*, 2019.
- [91] S. Langner, F. Häse, J.D. Perea, T. Stubhan, J. Hauch, L.M. Roch, T. Heumueller, A. Aspuru-Guzik, and C.J. Brabec. Beyond ternary OPV: High-throughput experimentation and self-driving laboratories optimize multi-component systems. *Adv. Mater.*, 32:1907801, 2020.

- [92] H. Park, N. Heldman, P. Rebentrost, L. Abbondanza, A. Iagatti, A. Alessi, B. Patrizi, M. Salvalaggio, L. Bussotti, M. Mohseni, F. Caruso, H.C. Johnsen, R. Fusco, P. Foggi, P.F. Scudo, S. Lloyd, and A. Belcher. Enhanced energy transport in genetically engineered excitonic networks. *Nat. Mater.*, 15(2):211, 2016.
- [93] Y. Ueda, H. Takeda, T. Yui, K. Koike, Y. Goto, S. Inagaki, and O. Ishitani. A visible-light harvesting system for CO₂ reduction using a RuII–ReI photocatalyst adsorbed in mesoporous organosilica. 8(3):439–442, 2015.
- [94] B. Qiu, Q. Zhu, M. Du, L. Fan, M. Xing, and J. Zhang. Efficient solar light harvesting CdS/Co₉S₈ hollow cubes for Z-scheme photocatalytic water splitting. *Angew. Chem. Int. Ed.*, 56(10):2684–2688, 2017.
- [95] J.M. Olson. Photosynthesis in the archean era. *Photosynth. Res.*, 88(2):109–117, 2006.
- [96] R.E. Blankenship, D.M. Tiede, J. Barber, G.W. Brudvig, G. Fleming, M. Ghirardi, M.R. Gunner, W. Junge, D.M. Kramer, A. Melis, T.A. Moore, C.C. Moser, D.G. Nocera, A.J. Nozik, D.R. Ort, W.W. Parson, R.C. Prince, and R.T. Sayre. Comparing photosynthetic and photovoltaic efficiencies and recognizing the potential for improvement. *Science*, 332(6031):805–809, 2011.
- [97] G.S. Schlau-Cohen. Principles of light harvesting from single photosynthetic complexes. *Interface Focus*, 5(3):20140088, 2015.
- [98] C.A. Wraight and R.K. Clayton. The absolute quantum efficiency of bacteriochlorophyll photooxidation in reaction centres of rhodospseudomonas spheroides. *Biochim. Biophys. Acta, Bioenerg.*, 333(2):246–260, 1974.
- [99] X.G. Zhu, S.P. Long, and D.R. Ort. Improving photosynthetic efficiency for greater yield. *Annu. Rev. Plant. Biol.*, 61:235–261, 2010.
- [100] D.A. Walker. Biofuels, facts, fantasy, and feasibility. *J. Appl. Phycol.*, 21(5):509–517, 2009.
- [101] R.H. Wijffels and M.J. Barbosa. An outlook on microalgal biofuels. *Science*, 329(5993):796–799, 2010.
- [102] S.H. Park, A. Roy, S. Beaupré, S. Cho, N. Coates, J.S. Moon, D. Moses, M. Leclerc, K. Lee, and A.J. Heeger. Bulk heterojunction solar cells with internal quantum efficiency approaching 100%. *Nat. Photon.*, 3(5):297, 2009.
- [103] S.C. Davis, K.J. Anderson-Teixeira, and E.H. DeLucia. Life-cycle analysis and the ecology of biofuels. *Trends Plant Sci.*, 14(3):140–146, 2009.
- [104] A.F. Sherwani and J.A. Usmani. Life cycle assessment of solar PV based electricity generation systems: A review. *Renew. Sust. Energy. Rev.*, 14(1):540–544, 2010.
- [105] B. Demmig-Adams and W.W. Adams III. Photoprotection in an ecological context: the remarkable complexity of thermal energy dissipation. *New Phytol.*, 172(1):11–21, 2006.

-
- [106] A.A. Pascal, Z. Liu, K. Broess, B. van Oort, H. van Amerongen, C. Wang, P. Horton, B. Robert, W. Chang, and A. Ruban. Molecular basis of photoprotection and control of photosynthetic light-harvesting. *Nature*, 436(7047):134, 2005.
- [107] S. Valteau. *Theoretical study of exciton transport in natural and synthetic light-harvesting systems*. PhD thesis, 2016.
- [108] D. Gülen. Interpretation of the excited-state structure of the Fenna-Matthews-Olson pigment protein complex of *Prosthecochloris Aestuarii* based on the simultaneous simulation of the 4K absorption, linear dichroism, and singlet-triplet absorption difference spectra: A possible excitonic explanation? *J. Phys. Chem.*, 100(44):17683–17689, 1996.
- [109] S. Savikhin, D.R. Buck, and W.S. Struve. Pump-probe anisotropies of fenna-matthews-olson protein trimers from *chlorobium tepidum*: a diagnostic for exciton localization? *Biophys. J.*, 73(4):2090–2096, 1997.
- [110] M. Wendling, T. Pullerits, M.A. Przyjalowski, S.I.E. Vulto, T.J. Aartsma, R. van Grondelle, and H. van Amerongen. Electron-vibrational coupling in the Fenna-Matthews-Olson complex of *Prosthecochloris Aestuarii* determined by temperature-dependent absorption and fluorescence line-narrowing measurements. *J. Phys. Chem. B*, 104(24):5825–5831, 2000.
- [111] J. Adolphs and T. Renger. How proteins trigger excitation energy transfer in the FMO complex of green sulfur bacteria. *Biophys. J.*, 91(8):2778–2797, 2006.
- [112] F. Müh, M.E.A. Madjet, J. Adolphs, A. Abdurahman, B. Rabenstein, H. Ishikita, E.W. Knapp, and T. Renger. α -helices direct excitation energy flow in the fenna-matthews-olson protein. *Proc. Natl. Acad. Sci.*, 104(43):16862–16867, 2007.
- [113] E.L. Read, G.S. Schlau-Cohen, G.S. Engel, J. Wen, R.E. Blankenship, and G.R. Fleming. Visualization of excitonic structure in the Fenna-Matthews-Olson photosynthetic complex by polarization-dependent two-dimensional electronic spectroscopy. 95(2):847–856, 2008.
- [114] S. Shim, P. Rebentrost, S. Valteau, and A. Aspuru-Guzik. Atomistic study of the long-lived quantum coherences in the Fenna-Matthews-Olson complex. *Biophys. J.*, 102(3):649–660, 2012.
- [115] G.S. Engel, T.R. Calhoun, E.L. Read, T.K. Ahn, T. Mancal, Y.C. Cheng, R.E. Blankenship, and G.R. Fleming. Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature*, 446:782–786, 2007.
- [116] A. Olaya-Castro, C.F. Lee, F.F. Olsen, and N.F. Johnson. Efficiency of energy transfer in a light-harvesting system under quantum coherence. *Phys. Rev. B*, 78(8):085115, 2008.
- [117] A. Ishizaki and G.R. Fleming. Theoretical examination of quantum coherence in a photosynthetic system at physiological temperature. *Proc. Natl. Acad. Sci.*, 106(41):17255–17260, 2009.

- [118] A. Ishizaki and G.R. Fleming. Unified treatment of quantum coherent and incoherent hopping dynamics in electronic energy transfer: Reduced hierarchy equation approach. *J. Chem. Phys.*, 130:234111, 2009.
- [119] A. Ishizaki and G.R. Fleming. On the adequacy of the Redfield equation and related approaches to the study of quantum dynamics in electronic energy transfer. *J. Chem. Phys.*, 130(23):234110, 2009.
- [120] P. Rebentrost, M. Mohseni, and A. Aspuru-Guzik. Role of quantum coherence and environmental fluctuations in chromophoric energy transport. *J. Phys. Chem. B*, 113(29):9942–9947, 2009.
- [121] F. Fassioli and A. Olaya-Castro. Distribution of entanglement in light-harvesting complexes and their quantum efficiency. *New J. Phys.*, 12:085006, 2010.
- [122] C. Olbrich, T.L.C. Jansen, J. Liebers, M. Aghtar, J. Strümpfer, K. Schulten, J. Knoester, and U. Kleinekathöfer. From atomistic modeling to excitation transfer and two-dimensional spectra of the FMO light-harvesting complex. *J. Phys. Chem. B*, 115(26):8609–8621, 2011.
- [123] L.A. Pachón and P. Brumer. Physical basis for long-lived electronic coherence in photosynthetic light-harvesting systems. *J. Phys. Chem. Lett.*, 2(21):2728–2732, 2011.
- [124] A. Ishizaki and G.R. Fleming. On the interpretation of quantum coherent beats observed in two-dimensional electronic spectra of photosynthetic light harvesting complexes. *J. Phys. Chem. B*, 115(19):6227–6233, 2011.
- [125] J. Yuen-Zhou, J.J. Krich, M. Mohseni, and A. Aspuru-Guzik. Quantum state and process tomography of energy transfer systems via ultrafast spectroscopy. *Proc. Natl. Acad. Sci.*, 108(43):17615–17620, 2011.
- [126] I. Kassal, J. Yuen-Zhou, and S. Rahimi-Keshari. Does coherence enhance transport in photosynthesis? *J. Phys. Chem. Lett.*, 4(3):362–367, 2013.
- [127] G.T. Oostergetel, H. van Amerongen, and E.J. Boekema. The chlorosome: A prototype for efficient light harvesting in photosynthesis. *Photosynth. Res.*, 104(2-3):245–255, 2010.
- [128] M.O. Pedersen, J. Linnanto, N.U. Frigaard, N.C. Nielsen, and M. Miller. A model of the protein–pigment baseplate complex in chlorosomes of photosynthetic green bacteria. *Photosynth. Res.*, 104(2-3):233–243, 2010.
- [129] R.E. Fenna and B.W. Matthews. Chlorophyll arrangement in a bacteriochlorophyll protein from *Chlorobium limicola*. *Nature*, 258(5536):573, 1975.
- [130] R.E. Fenna, B.W. Matthews, J.M. Olson, and E.K. Shaw. Structure of a bacteriochlorophyll-protein from the green photosynthetic bacterium *Chlorobium limicola*: crystallographic evidence for a trimer. *J. Mol. Biol.*, 84(2):231–240, 1974.

-
- [131] S. Valteau, R.A. Studer, F. Häse, C. Kreisbeck, R.G. Saer, R.E. Blankenship, E.I. Shakhnovich, and A. Aspuru-Guzik. Absence of selection for quantum coherence in the Fenna–Matthews–Olson complex: A combined evolutionary and excitonic study. *ACS Cent. Sci.*, 3(10):1086–1095, 2017.
- [132] M. Sarovar, A. Ishizaki, G.R. Fleming, and K.B. Whaley. Quantum entanglement in photosynthetic light-harvesting complexes. 6(6):462, 2010.
- [133] E. Brunk and U. Rothlisberger. Mixed quantum mechanical/molecular mechanical molecular dynamics simulations of biological systems in ground and electronically excited states. 115(12):6217–6263, 2015.
- [134] V.S. Pande, I. Baker, J. Chapman, S.P. Elmer, S. Khaliq, S.M. Larson, Y.M. Rhee, M.R. Shirts, C.D. Snow, E.J. Sorin, and B. Zagrovic. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68(1):91–109, 2003.
- [135] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103(2):227–249, 1976.
- [136] M. Aghtar, J. Strümpfer, C. Olbrich, K. Schulten, and U. Kleinekathöfer. Different types of vibrations interacting with electronic excitations in phycoerythrin 545 and Fenna–Matthews–Olson antenna systems. *J. Phys. Chem. Lett.*, 5(18):3131–3137, 2014.
- [137] D.E. Chandler, J. Strümpfer, M. Sener, S. Scheuring, and K. Schulten. Light harvesting by lamellar chromatophores in rhodospirillum photometricum. 106(11):2503–2510, 2014.
- [138] M.K. Lee and D.F. Coker. Modeling electronic-nuclear interactions for excitation energy transfer processes in light-harvesting complexes. *J. Phys. Chem. Lett.*, 7(16):3171–3178, 2016.
- [139] S.M. Blau, D.I.G. Bennett, C. Kreisbeck, G.D. Scholes, and A. Aspuru-Guzik. Local protein solvation drives direct down-conversion in phycobiliprotein PC645 via incoherent vibronic transport. *Proc. Natl. Acad. Sci.*, 115(15):E3342–E3350, 2018.
- [140] Y. Tanimura. Reduced hierarchy equations of motion approach with Drude plus Brownian spectral distribution: Probing electron transfer processes by means of two-dimensional correlation spectroscopy. *J. Chem. Phys.*, 137(22):22A550, 2012.
- [141] Y. Tanimura and R. Kubo. Time evolution of a quantum system in contact with a nearly Gaussian-Markoffian noise bath. *J. Phys. Soc. Jap.*, 58(1):101–114, 1989.
- [142] R. Hartmann and W.T. Strunz. Exact open quantum system dynamics using the hierarchy of pure states (HOPS). *J. Chem. Theory Comput.*, 13(12):5834–5845, 2017.
- [143] D. Suess, A. Eisfeld, and W. T. Strunz. Hierarchy of stochastic pure states for open quantum system dynamics. *Phys. Rev. Lett.*, 113:150403, 2014.

- [144] H.P. Breuer and F. Petruccione. *The theory of open quantum systems*. Oxford University Press on Demand, 2002.
- [145] J. Sun, J. Zhang, M. Zhang, M. Antonietti, X. Fu, and X. Wang. Bioinspired hollow semiconductor nanospheres as photosynthetic nanoparticles. *Neural Comput.*, 3:1139, 2012.
- [146] M.A. Green, E.D. Dunlop, D.H. Levi, J. Hohl-Ebinger, M. Yoshita, and A.W.Y. Ho-Baillie. Solar cell efficiency tables (version 54). *Prog. Photovolt.*, 27(NREL/JA-5K00-74116), 2019.
- [147] J. Britt and C. Ferekides. Thin-film cds/cdte solar cell with 15.8% efficiency. *Appl. Phys. Lett.*, 62(22):2851–2852, 1993.
- [148] J.M. Burst, J.N. Duenow, D.S. Albin, E. Colegrove, M.O. Reese, J.A. Aguiar, C.S. Jiang, M.K. Patel, M.M. Al-Jassim, D. Kuciauskas, S. Swain, T. Ablekim, K.G. Lynn, and W.K. Metzger. CdTe solar cells with open-circuit voltage breaking the 1 V barrier. *Nat. Energy*, 1(3):16015, 2016.
- [149] R. Kamada, T. Yagioka, S. Adachi, A. Handa, K.F. Tai, T. Kato, and H. Sugimoto. New world record Cu(In,Ga)(Se,S)₂ thin film solar cell efficiency beyond 22%. In *2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC)*, pages 1287–1291. IEEE, 2016.
- [150] H.J. Snaith. Perovskites: the emergence of a new era for low-cost, high-efficiency solar cells. *J. Phys. Chem. Lett.*, 4(21):3623–3630, 2013.
- [151] N.J. Jeon, J.H. Noh, W.S. Yang, Y.C. Kim, S. Ryu, J. Seo, and S.I. Seok. Compositional engineering of perovskite materials for high-performance solar cells. *Nature*, 517(7535):476, 2015.
- [152] W.S. Yang, J.H. Noh, N.J. Jeon, Y.C. Kim, S. Ryu, J. Seo, and S.I. Seok. High-performance photovoltaic perovskite layers fabricated through intramolecular exchange. *Science*, 348(6240):1234–1237, 2015.
- [153] W. Nie, H. Tsai, R. Asadpour, J.C. Blancon, A.J. Neukirch, G. Gupta, J.J. Crochet, M. Chhowalla, S. Tretiak, M.A. Alam, H.L. Wang, and A.D. Mohite. High-efficiency solution-processed perovskite solar cells with millimeter-scale grains. *Science*, 347(6221):522–525, 2015.
- [154] H. Tan, A. Jain, O. Voznyy, X. Lan, F.P.G. De Arquer, J.Z. Fan, R. Quintero-Bermudez, M. Yuan, B. Zhang, Y. Zhao, F. Fan, P. Li, L.N. Quan, Y. Zhao, Z.H. Lu, Z. Yang, S. Hoogland, and E.H. Sargent. Efficient and stable solution-processed planar perovskite solar cells via contact passivation. *Science*, 355(6326):722–726, 2017.
- [155] N. Gasparini, A. Salleo, I. McCulloch, and D. Baran. The role of the third component in ternary organic solar cells. page 1, 2019.
- [156] C. Yan, S. Barlow, Z. Wang, H. Yan, A.K.Y. Jen, S.R. Marder, and X. Zhan. Non-fullerene acceptors for organic solar cells. *Nat. Rev. Mater.*, 3(3):18003, 2018.

-
- [157] J. Hou, O. Inganäs, R.H. Friend, and F. Gao. Organic solar cells based on non-fullerene acceptors. *Nat. Mater.*, 17(2):119, 2018.
- [158] T. Moench, P. Friederich, F. Holzmueller, B. Rutkowski, J. Benduhn, T. Strunk, C. Koenner, K. Vandewal, A. Czyska-Filemonowicz, W. Wenzel, and K. Leo. Influence of meso and nanoscale structure on the properties of highly efficient small molecule solar cells. *Adv. Energy Mater.*, 6(4):1501280, 2016.
- [159] D. Venkateshvaran, M. Nikolka, A. Sadhanala, V. Lemaur, M. Zelazny, M. Kepa, M. Hurhangee, A.J. Kronemeijer, V. Pecunia, I. Nasrallah, I. Romanov, K. Broch, I. McCulloch, D. Emin, Y. Olivier, J. Cornil, D. Beljonne, and H. Sirringhaus. Approaching disorder-free transport in high-mobility conjugated polymers. *Nature*, 515(7527):384, 2014.
- [160] J. Zhang, H.S. Tan, X. Guo, A. Facchetti, and H. Yan. Material insights and challenges for non-fullerene organic solar cells based on small molecular acceptors. *Nat. Energy*, 3(9):720, 2018.
- [161] Y. He and Y. Li. Fullerene derivative acceptors for high performance polymer solar cells. *Phys. Chem. Chem. Phys.*, 13(6):1970–1983, 2011.
- [162] Y. Liu, J. Zhao, Z. Li, C. Mu, W. Ma, H. Hu, K. Jiang, H. Lin, H. Ade, and H. Yan. Aggregation and morphology control enables multiple cases of high-efficiency polymer solar cells. *Nat. Commun.*, 5:5293, 2014.
- [163] J. Zhao, Y. Li, G. Yang, K. Jiang, H. Lin, H. Ade, W. Ma, and H. Yan. Efficient organic solar cells processed from hydrocarbon solvents. *Nat. Energy*, 1(2):15027, 2016.
- [164] Y. He, H.Y. Chen, J. Hou, and Y. Li. Indene- C_{60} bisadduct: a new acceptor for high-performance polymer solar cells. *J. Am. Chem. Soc.*, 132(4):1377–1382, 2010.
- [165] P. Cheng and X. Zhan. Stability of organic solar cells: challenges and strategies. *Chem. Soc. Rev.*, 45(9):2544–2582, 2016.
- [166] S. Holliday, R.S. Ashraf, A. Wadsworth, D. Baran, S.A. Yousaf, C.B. Nielsen, C.H. Tan, S.D. Dimitrov, Z. Shang, N. Gasparini, M. Alamoudi, D. Laquai, C.J. Brabec, A. Salleo, J.R. Durrant, and I. McCulloch. High-efficiency and air-stable p3ht-based polymer solar cells with a new non-fullerene acceptor. *Nat. Commun.*, 7:11585, 2016.
- [167] D. Sun, D. Meng, Y. Cai, B. Fan, Y. Li, W. Jiang, L. Huo, Y. Sun, and Z. Wang. Non-fullerene-acceptor-based bulk-heterojunction organic solar cells with efficiency over 7%. *J. Am. Chem. Soc.*, 137(34):11156–11162, 2015.
- [168] W. Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008.
- [169] A. Hájek. Interpretations of probability. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.

- [170] I. Hacking. *Logic of statistical inference*. Cambridge University Press, 2016.
- [171] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian data analysis*. CRC press, 2013.
- [172] C.M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [173] B. De Finetti. Logical foundations and measurement of subjective probability. *Acta Psychol. (Amst.)*, 1970.
- [174] R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.
- [175] A.N. Kolmogorov. Foundations of the theory of probability. *Chelsea Publishing Co.*, 1950.
- [176] A.N. Kolmogorov. Grundlagen der Wahrscheinlichkeitsrechnung. 1933.
- [177] S.M. Ross et al. *A first course in probability*, volume 7. Pearson Prentice Hall Upper Saddle River, NJ, 2006.
- [178] A. Downey. *Think Bayes: Bayesian statistics in python*. O’Reilly Media, Inc., 2013.
- [179] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. London*, 186(1007):453–461, 1946.
- [180] S.M. Senkan. High-throughput screening of solid-state catalyst libraries. *Nature*, 394(6691):350, 1998.
- [181] W.F. Maier, K. Stoewe, and S. Sieg. Combinatorial and high-throughput materials science. *Angew. Chem. Int. Ed.*, 46(32):6016–6067, 2007.
- [182] R. Macarron, M.N. Banks, D. Bojanic, D.J. Burns, D.A. Cirovic, T. Garyantes, D.V.S. Green, R.P. Hertzberg, W.P. Janzen, J.W. Paslay, U. Schopfer, and S. Sittampalam. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.*, 10(3):188, 2011.
- [183] R.L. Martin, C.M. Simon, C. Smit, and M. Haranczyk. In silico design of porous polymer networks: High-throughput screening for methane storage materials. *J. Am. Chem. Soc.*, 136(13):5006–5022, 2014.
- [184] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281–305, 2012.
- [185] N. Baba. Convergence of a random optimization method for constrained optimization problems. *J. Optim. Theory Appl.*, 33(4):451–461, 1981.
- [186] J. Matyas. Random optimization. 26(2):246–253, 1965.
- [187] I.M. Sobol. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.*, 7(4):784–802, 1967.

-
- [188] R.A. Fisher. *The design of experiments*. Oliver and Boyd; Edinburgh; London, 1937.
- [189] G.E.P. Box, J.S. Hunter, and W.G. Hunter. *Statistics for experimenters: Design, innovation and discovery*. Wiley, 2nd edition, 2005.
- [190] M.J. Anderson and P.J. Whitcomb. *DOE simplified: Practical tools for effective experimentation*. CRC Press, 2016.
- [191] B.J. Reizman and K.F. Jensen. Feedback in flow for accelerated reaction development. *Acc. Chem. Res.*, 49:1786, 2016.
- [192] B.E. Walker, J.H. Bannock, A.M. Nightingale, and J.C. deMello. Tuning reaction products by constrained optimisation. *React. Chem. Eng.*, 2(5):785–798, 2017.
- [193] L. Cheng, R.S. Assary, X. Qu, A. Jain, S.P. Ong, N.N. Rajput, K. Persson, and L.A. Curtiss. Accelerating electrolyte discovery for energy storage with high-throughput screening. *J. Phys. Chem. Lett.*, 6(2):283–291, 2015.
- [194] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.
- [195] M.R. Hestenes and E. Stiefel. volume 49. NBS Washington, DC, 1952.
- [196] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, New York, 1987.
- [197] C.G. Broyden. The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA J. Appl. Math.*, 6(1):76–90, 1970.
- [198] J.P. McMullen, M.T. Stone, S.L. Buchwald, and K.F. Jensen. An integrated microreactor system for self-optimization of a heck reaction: From micro-to mesoscale flow systems. *Angew. Chem. Int. Ed.*, 49(39):7076–7080, 2010.
- [199] J.H. Holland and D. Goldberg. Genetic algorithms in search, optimization and machine learning. *Massachusetts: Addison-Wesley*, 1989.
- [200] R. John. Genetic programming: On the programming of computers by means of natural selection, 1992.
- [201] M. Srinivas and L.M. Patnaik. Genetic algorithms: A survey. *Comput.*, 27(6):17–26, 1994.
- [202] S. Mirjalili, J.S. Dong, A.S. Sadiq, and H. Faris. Genetic algorithm: Theory, literature review, and application in image reconstruction. In *Nature-Inspired Optimizers*, pages 69–85. Springer, 2020.
- [203] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, and B. Maruyama. Autonomy in materials research: A case study in carbon nanotube growth. *npj Comput. Mater.*, 2:16031, 2016.
- [204] A. Shayeghi, D. Götz, J.B.A. Davis, R. Schäfer, and R.L. Johnston. Pool-BCGA: A parallelised generation-free genetic algorithm for the ab initio global optimisation of nanoalloy clusters. *Phys. Chem. Chem. Phys.*, 17(3):2104–2112, 2015.

- [205] M. Schneider, M. Wilke, M.L. Hebestreit, J.A. Ruiz-Santoyo, L. Álvarez-Valtierra, T.Y. John, W.L. Meerts, D.W. Pratt, and M. Schmitt. Rotationally resolved electronic spectroscopy of the rotamers of 1, 3-dimethoxybenzene. *Phys. Chem. Chem. Phys.*, 19(32):21364–21372, 2017.
- [206] N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evol. Comput.*, 11(1):1–18, 2003.
- [207] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evol. Comput.*, 9(2):159–195, 2001.
- [208] R. Eberhart and J. Kennedy. A new optimizer using particle swarm theory. In *Proc. Int. Symp. - Micro Mach. Hum. Sci.*, pages 39–43. IEEE, 1995.
- [209] Y. Shi and R. Eberhart. A modified particle swarm optimizer. In *IEEE World Cong. Comput. Intell. - Evol. Comput. Proc.*, pages 69–73. IEEE, 1998.
- [210] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [211] V. Černý. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *J. Optim. Theory Appl.*, 45(1):41–51, 1985.
- [212] F. Glover. Tabu search—part II. *ORSA J. Comput.*, 2(1):4–32, 1990.
- [213] F. Glover. Tabu search—part I. *ORSA J. Comput.*, 1(3):190–206, 1989.
- [214] F. Glover. Future paths for integer programming and links to artificial intelligence. 13(5):533–549, 1986.
- [215] J. Močkus. *Bayesian approach to global optimization: Theory and applications*, volume 37. Springer Science & Business Media, 2012.
- [216] J. Močkus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.
- [217] H. J. Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic. Eng.*, 86(1):97–106, 1964.
- [218] J. Močkus. On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404, 1975.
- [219] J. Močkus. The Bayesian approach to global optimization. *Sys. Model. Optim.*, pages 473–481, 1982.
- [220] M.A. Osborne, R. Garnett, and S.J. Roberts. Gaussian processes for global optimization. In *Internat. Conf. Learn. Intell. Optim.*, pages 1–15, 2009.
- [221] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, and N. De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE*, 104(1):148–175, 2015.

-
- [222] D.R. Jones. A taxonomy of global optimization methods based on response surfaces. *J. Glob. Opt.*, 21(4):345–383, 2001.
- [223] M. Feurer, A. Klein, K. Eggenberger, J. Springenberg, M. Blum, and F. Hutter. Methods for improving Bayesian optimization for AutoML. In *Internat. Conf. Mach. Learn.*, 2015.
- [224] C. Thornton, F. Hutter, H.H. Hoos, and K. Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855, 2013.
- [225] K. Swersky, J. Snoek, and R.P. Adams. Multi-task Bayesian optimization. In *Adv. Neural. Inf. Process. Syst.*, pages 2004–2012, 2013.
- [226] A. Forrester, A. Sobester, and A. Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.
- [227] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *J. Glob. Opt.*, 13(4):455–492, 1998.
- [228] J. von Kügelgen, P.K. Rubenstein, B. Schölkopf, and A. Weller. Optimal experimental design via Bayesian optimization: Active causal structure learning for Gaussian process networks. *arXiv preprint arXiv:1910.03962*, 2019.
- [229] A. Foster, M. Jankowiak, E. Bingham, P. Horsfall, Y.W. Teh, T. Rainforth, and N. Goodman. Variational Bayesian optimal experimental design. In *Adv. Neural. Inf. Process. Syst.*, pages 14036–14047, 2019.
- [230] D.M. Negoescu, P.I. Frazier, and W.B. Powell. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS J. Comput.*, 23(3):346–363, 2011.
- [231] J. Vanlier, C.A. Tiemann, P.A.J. Hilbers, and N.A.W. van Riel. A Bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142, 2012.
- [232] P.I. Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.
- [233] F. Häse, L.M. Roch, C. Kreisbeck, and A. Aspuru-Guzik. Phoenix: A Bayesian optimizer for chemistry. *ACS Cent. Sci.*, 4(9):1134–1145, 2018.
- [234] F. Häse, L.M. Roch, and A. Aspuru-Guzik. Gryffin: An algorithm for Bayesian optimization for categorical variables informed by physical intuition with applications to chemistry. *arXiv preprint arXiv:2003.12127*, 2020.
- [235] F. Häse, L.M. Roch, and A. Aspuru-Guzik. Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chem. Sci.*, 9(39):7642–7655, 2018.

- [236] R. Martinez-Cantin, N. de Freitas, E. Brochu, J. Castellano, and A. Doucet. A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Autonom. Robots*, 27:29–103, 2009.
- [237] J. Snoek, H. Larochelle, and R.P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Adv. Neural. Inf. Process. Syst.*, volume 25, pages 2951–2959. 2012.
- [238] T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization. *J. Mach. Learn. Res.*, 15(1):3873–3923, 2014.
- [239] C.E. Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.
- [240] L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, 2001.
- [241] F. Hutter, H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithmic configuration. volume 5. 2011.
- [242] F. Hutter, H. Hoos, and K. Leyton-Brown. Parallel algorithm configuration. *Learn. Intell. Optim.*, pages 55–70, 2012.
- [243] M. Lindauer, K. Eggenberger, M. Feurer, S. Falkner, A. Biedenkapp, and F. Hutter. SMAC v3: Algorithm configuration in Python. <https://github.com/automl/SMAC3>, 2017.
- [244] T.K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.
- [245] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- [246] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, Prabhat, and R.P. Adams. Scalable Bayesian optimization using deep neural networks. In *Internat. Conf. Mach. Learn.*, volume 37, pages 2171–2180, 2015.
- [247] J.R. Springenberg, A. Klein, S. Falkner, and F. Hutter. Bayesian optimization with robust Bayesian neural networks. In *Adv. Neural. Inf. Process. Syst.*, pages 4134–4142, 2016.
- [248] N. Scrinivas, A. Krause, S.M. Kakade, and M.W. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory*, 58:3250–3265, 2012.
- [249] N. Scrinivas, A. Krause, M. Seeger, and S.M. Kakade. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proc. Int. Conf. Mach. Learn.*, pages 1015–1022, 2010.

-
- [250] E. Contal, D. Buffoni, A. Robicquet, and N. Vayatis. Parallel Gaussian process optimization with upper confidence bound and pure exploration. In *Eur. Conf. Mach. Learn. Know. Discov. Databases*, pages 225–240, 2013.
- [251] J.M. Hernández-Lobato, M. Gelbart, M. Hoffman, R.P. Adams, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Adv. Neural. Inf. Process. Syst.*, pages 918–926, 2014.
- [252] J.M. Hernández-Lobato, M. Gelbart, M. Hoffman, R.P. Adams, and Z. Ghahramani. Predictive entropy search for Bayesian optimization with unknown constraints. In *Internat. Conf. Mach. Learn.*, volume 37, pages 1699–1707, 2015.
- [253] R.T. Marler and J.S. Arora. Survey of multi-objective optimization methods for engineering. *Struct. Multidiscipl. Optim.*, 26(6):369–395, 2004.
- [254] V. Pareto. *Manuale di economica politica*, societa editrice libraria. Milan. *translated to English by Schwier AS as Manual of Political Economy, Kelley, New York*, 1906.
- [255] K. Miettinen. *Nonlinear multiobjective optimization*, volume 12 of international series in operations research and management science, 1999.
- [256] I. Das and J.E. Dennis. Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM J. Optim.*, 8(3):631–657, 1998.
- [257] R. Motta, S.M.B. Afonso, and P.R.M. Lyra. A modified NBI and NC method for the solution of N-multiobjective optimization problems. *Struct. Multidiscipl. Optim.*, 46(2):239–259, 2012.
- [258] A. Messac, A. Ismail-Yahaya, and C.A. Mattson. The normalized normal constraint method for generating the Pareto frontier. *Struct. Multidiscipl. Optim.*, 25(2):86–98, 2003.
- [259] A. Messac and C.A. Mattson. Normal constraint method with guarantee of even representation of complete Pareto frontier. *AIAA J.*, 42(10):2101–2111, 2004.
- [260] D. Mueller-Gritschneider, H. Graeb, and U. Schlichtmann. A successive approach to compute the bounded Pareto front of practical multiobjective optimization problems. *SIAM J. Optim.*, 20(2):915–934, 2009.
- [261] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2):182–197, 2002.
- [262] D.V. Vargas, J. Murata, H. Takano, and A.C.B. Delbem. General subpopulation framework and taming the conflict inside populations. *Evol. Comput.*, 23(1):1–36, 2015.
- [263] D. Hernández-Lobato, J. Hernandez-Lobato, A. Shah, and R.P. Adams. Predictive entropy search for multi-objective Bayesian optimization. In *Internat. Conf. Mach. Learn.*, pages 1492–1501, 2016.

- [264] V. Picheny. Multiobjective optimization using Gaussian process emulators via stepwise uncertainty reduction. *Stat. Comput.*, 25(6):1265–1280, 2015.
- [265] M. Emmerich and J.W. Klinkenberg. The computation of the expected improvement in dominated hyper volume of Pareto front approximations. *Rapport technique, Leiden University*, 34, 2008.
- [266] W. Ponweiser, T. Wagner, D. Biermann, and M. Vincze. Multiobjective optimization on a limited budget of evaluations using model-assisted s-metric selection. In *International Conference on Parallel Problem Solving from Nature*, pages 784–794. Springer, 2008.
- [267] J. Knowles. ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Trans. Evol. Comput.*, 10(1):50–66, 2006.
- [268] M. Pescador-Rojas, R.H. Gómez, E. Montero, N. Rojas-Morales, M.C. Riff, and C.A. Coello. An overview of weighted and unconstrained scalarizing functions. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 499–513, 2017.
- [269] I.Y. Kim and O.L. de Weck. Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Struct. Multidiscipl. Optim.*, 29(2):149–158, 2005.
- [270] Y. Y. Haimes. On a bicriterion formulation of the problems of integrated system identification and system optimization. 1(3):296–297, 1971.
- [271] V. Changkong and Y.Y. Haimes. Multiobjective decision making: Theory and methodology. In *North-Holland Series in System Science and Engineering*, volume 8. Elsevier Science Publishing Co New York NY, 1983.
- [272] J.L. Cohon. *Multiobjective programming and planning*, volume 140. Courier Corporation, 2004.
- [273] C.L. Hwang and A.S.M. Masud. *Multiple objective decision making-methods and applications: A state-of-the-art survey*, volume 164. Springer Science & Business Media, 2012.
- [274] W. Stadler. Fundamentals of multicriteria optimization. In *Multicriteria Optimization in Engineering and in the Sciences*, pages 1–25. Springer, 1988.
- [275] O. Grodzevich and O. Romanko. Normalization and other topics in multi-objective optimization. *Proceedings of the Fields-MITACS Industrial Problems Workshop*, 2006.
- [276] F. Waltz. An engineering approach: Hierarchical optimization criteria. *IEEE Trans. Autom. Control*, 12(2):179–180, 1967.
- [277] M.J. Rentmeesters, W.K. Tsai, and K.J. Lin. A theory of lexicographic multi-criteria optimization. In *Second IEEE International Conference on Engineering of Complex Computer Systems, 1996. Proceedings.*, pages 76–79. IEEE, 1996.

-
- [278] S.J. Horvath, J.R. Firca, T. Hunkapiller, M.W. Hunkapiller, and L. Hood. An automated DNA synthesizer employing deoxynucleoside 3'-phosphoramidites. In *Methods in enzymology*, volume 154, pages 314–326. Elsevier, 1987.
- [279] R.B. Merrifield, J.M. Stewart, and N. Jernberg. Instrument for automated synthesis of peptides. *Anal. Chem.*, 38(13):1905–1914, 1966.
- [280] R.B. Merrifield. Automated synthesis of peptides. *Science*, 150(3693):178–185, 1965.
- [281] Otto Mayr. The origins of feedback control. *Sci. Am.*, 223(4):110–119, 1970.
- [282] J.L. Heilbron. *The Oxford companion to the history of modern science*. Oxford University Press, 2003.
- [283] M.C. Jackson. *Critical systems thinking and the management of complexity*, volume 1. John Wiley and Sons, 2019.
- [284] R.L. Hills. *Power from wind: a history of windmill technology*. Cambridge University Press, 1996.
- [285] S. Bennett. *A history of control engineering, 1930-1955*. Number 47. IET, 1993.
- [286] K. Wang, R. Liang, H. Chen, S. Lu, S. Jia, and W. Wang. A microfluidic immunoassay system on a centrifugal platform. *Sens. Actuators B Chem.*, 251:242–249, 2017.
- [287] Z. Cai, J. Xiang, B. Zhang, and W. Wang. A magnetically actuated valve for centrifugal microfluidic applications. *Sens. Actuators B Chem.*, 206:22–29, 2015.
- [288] Z. Cai, J. Xiang, and W. Wang. A pinch-valve for centrifugal microfluidic platforms and its application in sequential valving operation and plasma extraction. *Sens. Actuators B Chem.*, 221:257–264, 2015.
- [289] K. Olsen. The first 110 years of laboratory automation: Technologies, applications, and the creative scientist. *J. Lab. Autom.*, 17(6):469–480, 2012.
- [290] T.M. Stevens. Rapid and automatic filtration. *Am. Chemist.*, 6(3):102, 1875.
- [291] W.D. Horne. An automatic extractor. *J. Am. Chem. Soc.*, 15(5):270–272, 1893.
- [292] J.M. Pickel. An automatic filter washer. *J. Am. Chem. Soc.*, 23(8):589–593, 1901.
- [293] T.A. Mitchell. Automatic filter feed. *J. Ind. Eng. Chem*, 4:613, 1912.
- [294] E.C. Lathrop. A simple device for the automatic and intermittent washing of precipitates. *Ind. Eng. Chem.*, 9(5):527–528, 1917.
- [295] L.M. Dennis. *Gas analysis*. Macmillan, 1913.
- [296] G.B. Taylor and H.S. Taylor. Automatic volumetric analysis carbon monoxide recorder. *Ind. Eng. Chem.*, 14(11):1008–1009, 1922.
- [297] J.W. Frazer. Computer automation in chemistry. *Evol. Comput.*, 2(3):271–295, 1970.

- [298] B. Johnson, T. Kuga, and H.M. Gladney. Computer-assisted spectroscopy. *IBM J. Res. Dev.*, 13(1):36–45, 1969.
- [299] PM Grant. Automation of a wide-range, general-purpose spectrophotometric system. *IBM J. Res. Dev.*, 13(1):15–27, 1969.
- [300] H.M. Gladney. A simple time-sharing monitor system for laboratory automation. *J. Comp. Phys.*, 2(3):255–273, 1968.
- [301] J.N. Gayles, W.L. Honzik, and D.O. Wilson. On-line far-infrared michelson interferometry in a time-shared mode. *IBM J. Res. Dev.*, 14(1):25–32, 1970.
- [302] B.J. McGrattan and D.J. Macero. Computer program for laboratory robots. *Chem. Eng. News*, 62(48):37, 1984.
- [303] J.S. Lindsey. A retrospective on the automation of laboratory synthetic chemistry. *Chemom. Intell. Lab. Syst.*, 17(1):15–45, 1992.
- [304] H. Winicov, J. Schainbaum, J. Buckley, G. Longino, J. Hill, and C.E. Berkoff. Chemical process optimization by computer—a self-directed chemical synthesis system. *Anal. Chim. Acta*, 103(4):469–476, 1978.
- [305] T. Sugawara and D.G. Cork. Past and present development of automated synthesis apparatus for pharmaceutical chemistry at Takeda Chemical Industries. *Lab. Robotics Automat.*, 8:221–230, 1996.
- [306] C. Simms and J. Singh. Rapid process development and scale-up using a multiple reactor system. *Org. Proc. Res. Dev.*, 4:554–562, 2000.
- [307] Y.L. Dar. High-throughput experimentation: A powerful enabling technology for the chemicals and materials industry. *Macromol. Rapid Commun.*, 25:34–47, 2004.
- [308] T. Chapman. A structured approach. *Nature*, 421:661, 2003.
- [309] G. Schneider. Automating drug discovery. *Nat. Rev. Drug Discov.*, 17:97, 2018.
- [310] M. Sasaki, T. Kageoka, K. Ogura, H. Kataoka, T. Ueta, and S. Sugihara. Total laboratory automation in Japan: Past, present and the future. *Clin. Chim.*, 278(2):217–227, 1998.
- [311] R.A. Felder, J. Savory, K.S. Margrey, J.W. Holman, and J.C. Boyd. Development of a robotic near patient testing laboratory. *Arch. Pathol. Lab. Med.*, 119(10):948–951, 1995.
- [312] R.A. Felder. Clinical laboratory robotics in the 1990s. *Chemom. Intell. Lab. Syst.*, 17(1):111–118, 1992.
- [313] R.A. Felder, J.C. Boyd, J. Savory, K. Margrey, A. Martinez, and D. Vaughn. Robotics in the clinical laboratory. *Clin. Lab. Med.*, 8(4):699–712, 1988.

-
- [314] M. Nettekoven and A.W. Thomas. Accelerating drug discovery by integrative implementation of laboratory automation in the work flow. *Curr. Med. Chem.*, 9(23):2179–2190, 2002.
- [315] Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, 2004.
- [316] J. Boyd. Robotic laboratory automation. *Science*, 295(5554):517–518, 2002.
- [317] W.J. Coates, D.J. Hunter, and W.S. MacLachlan. Successful implementation of automation in medicinal chemistry. *Drug Discov.*, 5(11):521–527, 2000.
- [318] I. Clark-Lewis, R. Aebersold, H. Ziltener, J.W. Schrader, L.E. Hood, and S.B. Kent. Automated chemical synthesis of a protein growth factor for hemopoietic cells, interleukin-3. 231(4734):134–139, 1986.
- [319] E.M. Woerly, J. Roy, and M.D. Burke. Synthesis of most polyene natural product motifs using just 12 building blocks and one coupling reaction. *Nat. Chem.*, 6:484, 2014.
- [320] R.F. Service. The Synthesis Machine. *Science*, 347:1190, 2015.
- [321] J. Li, S.G. Ballmer, E.P. Gillis, S. Fujii, M.J. Schmidt, A.M.E. Palazzolo, J.W. Lehmann, G.F. Morehouse, and M.D. Burke. Synthesis of many different types of organic small molecules using one automated process. *Science*, (6227):1221–1226, 2015.
- [322] M.A.R. Meier, J.F. Gohy, C.A. Fustin, and U.S. Schubert. Combinatorial synthesis of star-shaped block copolymers: Host-guest chemistry of unimolecular reversed micelles. *J. Am. Chem. Soc.*, 126:11517–11521, 2004.
- [323] R. Hoogenboom, F. Wiesbrock, M.A.M. Leenen, M.A.R. Meier, and U.S. Schubert. Accelerating the living polymerization of 2-nonyl-2-oxazoline by implementing a microwave synthesizer into a high-throughput experimentation workflow. *J. Comb. Chem.*, 7:10–13, 2005.
- [324] A.J. Mijalis, D.A. Thomas III, M.D. Simon, A. Adamo, R. Beaumont, K.F. Jensen, and B.L. Pentelute. A fully automated flow-based approach for accelerated peptide synthesis. *Nat. Chem. Biol.*, 13:464, 2017.
- [325] D.C. Patel, Y.F. Lyu, J. Gandarilla, and S. Doherty. Unattended reaction monitoring using an automated microfluidic sampler and on-line liquid chromatography. *Anal. Chim. Acta*, 1004:32–39, 2018.
- [326] S. Chen, Y. Hou, H. Chen, X. Tang, S. Langner, N. Li, T. Stubhan, I. Levchuk, E. Gu, A. Osvet, and C.J. Brabec. Exploring the stability of novel wide bandgap perovskites by a robot based high throughput approach. *Adv. Energy Mater.*, 8(6):1701543, 2018.
- [327] E.M. Chan, C. Xu, A.W. Mao, G. Han, J.S. Owen, B.E. Cohen, and D.J. Milliron. Reproducible, high-throughput synthesis of colloidal nanocrystals for optimization in multidimensional parameter space. *Nano Lett.*, 10:1874–1885, 2010.

- [328] P. Raccuglia, K.C. Elbert, P.D.F. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier, and A.J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73, 2016.
- [329] J.Y. Pan. Engineering chemistry innovation. *ACS Med. Chem. Lett.*, 10(5):703–707, 2019.
- [330] F. Häse, C. Kreisbeck, and A. Aspuru-Guzik. Machine learning for quantum dynamics: Deep learning of excitation energy transfer properties. *Chem. Sci.*, 8(12):8419–8426, 2017.
- [331] S. Caffarri, R. Kouřil, S. Kereïche, E.J. Boekema, and R. Croce. Functional architecture of higher plant photosystem II supercomplexes. *EMBO J.*, 28(19):3052–3063, 2009.
- [332] N.R. Baker. Chlorophyll fluorescence: A probe of photosynthesis in vivo. *Annu. Rev. Plant. Biol.*, 59:89–113, 2008.
- [333] C. Kreisbeck and A. Aspuru-Guzik. Efficiency of energy funneling in the photosystem II supercomplex of higher plants. *Chem. Sci.*, 7:4174–4183, 2016.
- [334] K. Amarnath, D.I.G. Bennett, A.R. Schneider, and G.R. Fleming. Multiscale model of light harvesting by photosystem II in plants. *Proc. Natl. Acad. Sci.*, 113:1156–1161, 2016.
- [335] G.D. Scholes, G.R. Fleming, L.X. Chen, A. Aspuru-Guzik, A. Buchleitner, D.F. Coker, G.S. Engel, R. Van Grondelle, A. Ishizaki, D. M. Jonas, J.S. Lundeen, J.K. McCusker, S. Mukamel, J.P. Ogilvie, A. Olaya-Castro, M.A. Ratner, F.C. Spano, B. Whaley, and X. Zhu. Using coherence to enhance function in chemical and biophysical systems. *Nature*, 543(7647):647, 2017.
- [336] G. D. Scholes, G. R. Fleming, A. Olaya-Castro, and R. Van Grondelle. Lessons from nature about solar light harvesting. *Nat. Chem.*, 3(10):763, 2011.
- [337] C. Kreisbeck and T. Kramer. Long-lived electronic coherence in dissipative exciton dynamics of light-harvesting complexes. *J. Phys. Chem. Lett.*, 3(19):2828–2833, 2012.
- [338] A.W. Chin, J. Prior, R. Rosenbach, F. Caycedo-Soler, S.F. Huelga, and M.B. Plenio. The role of non-equilibrium vibrational structures in electronic coherence and recoherence in pigment-protein complexes. *Nat. Phys.*, 9:113–118, 2013.
- [339] N. Christensson, H.F. Kauffmann, T. Pullerits, and T. Mančal. Origin of long lived coherences in light-harvesting complexes. *J. Phys. Chem. B*, 116:7449–7454, 2012.
- [340] J.C. Dean, T. Mirkovic, Z.S.D. Toa, D.G. Oblinsky, and G.D. Scholes. Vibronic enhancement of algae light harvesting. *Chem*, 1:858–872, 2016.
- [341] E. Romero, R. Augulis, V.I. Novoderezhkin, M. Ferretti, J. Thieme, D. Zigmantas, and R. van Grondelle. Quantum coherence in photosynthesis for efficient solar-energy conversion. *Nat. Phys.*, 10:676–682, 2014.

-
- [342] A. De Sio, F. Troiani, M. Maiuri, J. Réhault, E. Sommer, J. Lim, S.F. Huelga, M.B. Plenio, C.A. Rozzi, G. Cerullo, E. Molinari, and C. Lienau. Vibronic origin of long-lived coherence in an artificial molecular light harvester. *Nat. Commun.*, 7:13742, 2016.
- [343] E. Collini, C.Y. Wong, K.E. Wilk, P.M.G. Curmi, P. Brumer, and G.D. Scholes. Coherently wired light-harvesting in photosynthetic marine algae at ambient temperature. *Nature*, 463:644–647, 2010.
- [344] T. Brixner, J. Stenger, H.M. Vaswani, M. Cho, R.E. Blankenship, and G.R. Fleming. Two-dimensional spectroscopy of electronic couplings in photosynthesis. *Nature*, 434:625–628, 2005.
- [345] J. Schulze and O. Kühn. Explicit correlated exciton-vibrational dynamics of the FMO complex. *J. Phys. Chem. B*, 119:6211–6216, 2015.
- [346] B. Hein, C. Kreisbeck, T. Kramer, and M. Rodríguez. Modelling of oscillations in two-dimensional echo-spectra of the Fenna-Matthews-Olson complex. *New J. Phys.*, 14:023018, 2012.
- [347] P. Nalbach, D. Braun, and M. Thorwart. Exciton transfer dynamics and quantumness of energy transfer in the Fenna-Matthews-Olson complex. *Phys. Rev. E*, 84:041926, 2011.
- [348] T. Scholak, T. Wellens, and A. Buchleitner. Optimal networks for excitonic energy transport. *J. Phys. B*, 44:184012, 2011.
- [349] S. Mostarda, F. Levi, D. Prada-Gravia, F. Mintert, and F. Rao. Structure-dynamics relationship in coherent transport through disordered systems. *Nat. Commun.*, 4:2296, 2013.
- [350] S. Baghbanzadeh and I. Kassal. Distinguishing the roles of energy funnelling and delocalization in photosynthetic light harvesting. *Phys. Chem. Chem. Phys.*, 18:7459–7467, 2016.
- [351] S. Baghbanzadeh and I. Kassal. Geometry, supertransfer and optimality in the light harvesting of purple bacteria. *J. Phys. Chem. Lett.*, 7:3804–3811, 2016.
- [352] G.C. Knee, P. Rowe, L.D. Smith, A. Troisi, and A. Datta. Structure-dynamics relation in physically-plausible multi-chromophore systems. *J. Phys. Chem. Lett.*, 8:2328–2333, 2017.
- [353] J.A. Leegwater. Coherent versus incoherent energy transfer and trapping in photosynthetic antenna complexes. *J. Phys. Chem.*, 100(34):14403–14409, 1996.
- [354] V. May. *Charge and energy transfer dynamics in molecular systems*. John Wiley & Sons, 2008.
- [355] F. Müh, M.E.A. Madjet, and T. Renger. Structure-based simulation of linear optical spectra of the CP43 core antenna of photosystem II. *Photosynth. Res.*, 111(1-2):87–101, 2012.

- [356] G. Raszewski and T. Renger. Light harvesting in photosystem II core complexes is limited by the transfer to the trap: Can the core complex turn into a photoprotective mode? *J. Am. Chem. Soc.*, 130(13):4431–4446, 2008.
- [357] G. Raszewski, B.A. Diner, E. Schlodder, and T. Renger. Spectroscopic properties of reaction center pigments in photosystem II core complexes: Revision of the multimer model. *Biophys. J.*, 95(1):105–119, 2008.
- [358] M. Mohseni, P. Rebentrost, S. Lloyd, and A. Aspuru-Guzik. Environment-assisted quantum walks in photosynthetic energy transfer. *J. Chem. Phys.*, 129:174106, 2008.
- [359] C. Kreisbeck, T. Kramer, M. Rodríguez, and B. Hein. High-performance solution of hierarchical equations of motion for studying energy transfer in light-harvesting complexes. *J. Chem. Theory Comput.*, 7(7):2166–2174, 2011.
- [360] P. Rebentrost, M. Mohseni, I. Kassal, S. Lloyd, and A. Aspuru-Guzik. Environment-assisted quantum transport. *New J. Phys.*, 11:033003, 2009.
- [361] F. Caruso, A.W. Chin, A. Datta, S.F. Huelga, and M.B. Plenio. Highly efficient energy excitation transfer in light-harvesting complexes: the fundamental role of noise-assisted transport. *J. Chem. Phys.*, 131:105106, 2009.
- [362] V. May and O. Kühn. *Charge and Energy Transfer Dynamics in Molecular Systems*. Wiley-VCH, Weinheim, 2004.
- [363] Y. C. Cheng and G. R. Fleming. Dynamics of light harvesting in photosynthesis. *Annu. Rev. Phys. Chem.*, 60:241–262, 2009.
- [364] Y. Yan, F. Yang, Y. Liu, and J. Shao. Hierarchical approach based on stochastic decoupling to dissipative systems. *Chem. Phys. Lett.*, 395:216–221, 2004.
- [365] R. Xu, P. Cui, C. Li, Y. Mo, and Y. Yan. Exact quantum master equation via the calculus on path integrals. *J. Chem. Phys.*, 112:041103, 2005.
- [366] A. Ishizaki and Y. Tanimura. Quantum dynamics of system strongly coupled to low-temperature colored noise bath: Reduced hierarchy equations approach. *J. Phys. Soc. Jap.*, 74:3131–3134, 2005.
- [367] C. Kreisbeck, T. Kramer, and A. Aspuru-Guzik. Scalable high-performance algorithm for the simulation of exciton dynamics. Application to the light-harvesting complex II in the presence of resonant vibrational modes. *J. Chem. Theory Comput.*, 10:4045–4054, 2014.
- [368] V. Novoderezhkin, A. Marin, and R. van Grondelle. Intra- and inter-monomeric transfers in the light harvesting LHCI complex: The Redfield-Förster picture. *Phys. Chem. Chem. Phys.*, 13:17093–17103, 2011.
- [369] Z.W. Ulissi, A.J. Medford, R. Bligaard, and J.K. Nørskov. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.*, 8:14621, 2017.

- [370] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O.A. von Lilienfeld, A. Tkatchenko, and K.R. Müller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.*, 9(8):3404–3419, 2013.
- [371] J. Behler and M. Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, 2007.
- [372] J.S. Smith, O. Isayev, and A.E. Roitberg. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.*, 8:3192–3203, 2017.
- [373] K. Yao, J. E. Herr, and J. Parkhill. The many-body expansion combined with neural networks. *J. Chem. Phys.*, 146:014106, 2017.
- [374] J. Gomes, B. Ramsundar, E. N. Feinberg, and V. Pande. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint arXiv:1703.10603*, 2017.
- [375] F. Häse, S. Valleau, E. Pyzer-Knapp, and A. Aspuru-Guzik. Machine learning exciton dynamics. *Chem. Sci.*, 7(8):5139–5147, 2016.
- [376] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.R. Müller, and O.A. von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.*, 15(9):095003, 2013.
- [377] F. Häse. Deep learning of excitation energy transfer properties at Redfield accuracy. <https://github.com/FlorianHase/LearningExcitonTransfer>, 2017.
- [378] A.R. Holzwarth, M.G. Müller, M. Reus, M. Nowaczyk, J. Sander, and M. Rögner. Kinetics and mechanism of electron transfer in intact photosystem II and in the isolated reaction center: pheophytin is the primary electron acceptor. *Proc. Natl. Acad. Sci.*, 103(18):6895–6900, 2006.
- [379] D. Kingma and J. Ba. Adam: A method for stochastic optimization. 2015.
- [380] L.C.W. Dixon and G.P. Szegö. *Towards global optimisation*. North-Holland Amsterdam, 1978.
- [381] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.
- [382] A.E. Jailaubekov, A.P. Willard, J.R. Tritsch, W.L. Chan, N. Sai, R. Gearba, L.G. Kaake, K.J. Williams, K. Leung, P.J. Rossky, and X.Y. Zhu. Hot charge-transfer

- excitons set the time limit for charge separation at donor/acceptor interfaces in organic photovoltaics. *Nat. Mater.*, 12(1):66, 2013.
- [383] D.A. Vithanage, A. Devižis, V. Abramavičius, Y. Infahsaeng, D. Abramavičius, R.C.I. MacKenzie, P.E. Keivanidis, A. Yartsev, D. Hertel, J. Nelson, V. Sundström, and V. Gulbinas. Visualizing charge separation in bulk heterojunction organic solar cells. *Nat. Commun.*, 4:2334, 2013.
- [384] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sanchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik. The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.*, 2:2241–2251, 2011.
- [385] L.M. Roch, S.K. Saikin, F. Häse, P. Friederich, R.H. Goldsmith, S. León, and A. Aspuru-Guzik. From absorption spectra to charge transfer in nanoaggregates of oligomers with machine learning. *ACS Nano*, 2019.
- [386] E. Cantatore, editor. *Applications of organic and printed electronics. A technology-enabled revolution*. Springer, 2013.
- [387] Y.J. Cheng, S.H. Yang, and C.S. Hsu. Synthesis of conjugated polymers for organic solar cell applications. *Chem. Rev.*, 109:5868–5923, 2009.
- [388] S. Logothetidis. Flexible organic electronic devices: Materials, process and applications. *Mater. Sci. Eng. B*, 152(1):96 – 104, 2008.
- [389] J. Rivnay, S. Inal, B.A. Collins, M. Sessolo, E. Stavrinidou, X. Strakosas, C. Tascone, D.M. Delongchamp, and G.G. Malliaras. Structural control of mixed ionic and electronic transport in conducting polymers. *Nat. Commun.*, 7:11287, 2016.
- [390] T. Lei, M. Guan, J. Liu, H.C. Lin, R. Pfattner, L. Shaw, A.F. McGuire, T.C. Huang, L. Shao, K.T. Cheng, J.B.H. Tok, and Z. Bao. Biocompatible and totally disintegrable semiconducting polymer for ultrathin and ultralightweight transient electronics. *Proc. Natl. Acad. Sci.*, 114(20):5107–5112, 2017.
- [391] J.Y. Oh, S. Rondeau-Gagné, Y.C. Chiu, A. Chortos, F. Lissel, G.J.N. Wang, B.C. Schroeder, T. Kurosawa, J. Lopez, T. Katsumata, J. Xu, C. Zhu, X. Gu, W.G. Bae, Y. Kim, L. Jin, J.W. Chung, J.B.H. Tok, and Z. Bao. Intrinsically stretchable and healable semiconducting polymer for organic transistors. *Nature*, 539:411, 2016.
- [392] M.U. Ocheje, B.P. Charron, A. Nyayachavadi, and S. Rondeau-Gagné. Stretchable electronics: Recent progress in the preparation of stretchable and self-healing semiconducting conjugated polymers. *Flexible Printed Electron.*, 2(4):043002, 2017.
- [393] A. Facchetti. π -conjugated polymers for organic electronics and photovoltaic cell applications. *Chem. Mater.*, 23(3):733–758, 2011.
- [394] D.J. Lipomi, B.C.K. Tee, M. Vosgueritchian, and Z. Bao. Stretchable organic solar cells. *Adv. Mater.*, 23(15):1771–1775, 2011.

-
- [395] Z. Guo, Y. Qiao, H. Liu, C. Ding, Y. Zhu, M. Wan, and L. Jiang. Self-assembled hierarchical micro/nano-structured PEDOT as an efficient oxygen reduction catalyst over a wide pH range. *J. Mater. Chem.*, 22(33):17153–17158, 2012.
- [396] B. Winther-Jensen, O. Winther-Jensen, M. Forsyth, and D.R. MacFarlane. High rates of oxygen reduction over a vapor phase-polymerized pedot electrode. *Science*, 321(5889):671–674, 2008.
- [397] N. Dubey and M. Leclerc. Conducting polymers: Efficient thermoelectric materials. *J. Polym. Sci. B*, 49(7):467–475, 2011.
- [398] H. Wang, U. Ail, R. Gabrielsson, M. Berggren, and X. Crispin. Ionic seebeck effect in conducting polymers. *Adv. Energy Mater.*, 5(11):1500044, 2015.
- [399] Q. Wei, M. Mukaida, K. Kirihara, Y. Naitoh, and T. Ishida. Recent progress on PEDOT-based thermoelectric materials. *Materials*, 8(2):732–750, 2015.
- [400] O. Bubnova, Z.U. Khan, A. Malti, S. Braun, M. Fahlman, M. Berggren, and X. Crispin. Optimization of the thermoelectric figure of merit in the conducting polymer poly(3,4-ethylenedioxythiophene). *Nat. Mater.*, 10(6):429–433, 2011.
- [401] K. Sun, S. Zhang, P. Li, Y. Xia, X. Zhang, D. Du, F.H. Isikgor, and J. Ouyang. Review on application of PEDOTs and PEDOT:PSS in energy conversion and storage devices. *J. Mater.*, 26(7):4438–4462, 2015.
- [402] L.M. Roch, L. Zoppi, J.S. Siegel, and K.K. Baldrige. Indenocorannulene-based materials: Effect of solid-state packing and intermolecular interactions on optoelectronic properties. *J. Phys. Chem. C*, 121(2):1220–1234, 2017.
- [403] D.A. Hinton, J.D. Ng, J. Sun, S. Lee, S.K. Saikin, J. Logsdon, D.S. White, A.N. Marquard, A.C. Cavell, V.K. Krasecki, K.A. Knapper, K.M. Lupo, M.R. Wasielewski, A. Aspuru-Guzik, J.S. Biteen, P. Gopalan, and R.H. Goldsmith. Mapping forbidden emission to structure in self-assembled organic nanoparticles. *J. Am. Chem. Soc.*, 140:15827–15841, 2018.
- [404] J. Luo, D. Billep, T. Waechtler, T. Otto, M. Toader, O. Gordan, E. Sheremet, J. Martin, M. Hietschold, D.R.T. Zahn, and T. Gessner. Enhancement of the thermoelectric properties of PEDOT:PSS thin films by post-treatment. *J. Mater. Chem. A*, 1:7576–7583, 2013.
- [405] J. Ouyang, Q. Xu, C.W. Chu, Y. Yang, G. Li, and J. Shinar. On the mechanism of conductivity enhancement in poly(3,4-ethylenedioxythiophene):poly(styrene sulfonate) film through solvent treatment. *Polymer*, 45:8443–8450, 2004.
- [406] S.K. Saikin, A. Eisfeld, S. Valteau, and A. Aspuru-Guzik. Photonics meets excitonics: natural and artificial molecular aggregates. *Nanophotonics*, 2:21–38, 2013.
- [407] L. Groenendaal, F. Jonas, D. Freitag, H. Pielartzik, and J R. Reynolds. Poly(3,4-ethylenedioxythiophene) and its derivatives: Past, present, and future. *Adv. Mater.*, 12(7):481–494, 2000.

- [408] W. Lövenich. PEDOT–Properties and Applications. *Polym. Sci. Ser. C*, 56:135–143, 2014.
- [409] A. Elschner, S. Kirchmeyer, W. Lovenich, U. Merker, and Reuter K. *PEDOT: Principles and applications of an intrinsically conductive polymer*. CRC Press, Taylor & Francis Group, 2011.
- [410] D.C. Martin, J. Wu, C.M. Shaw, Z. King, S.A. Spanninga, S. Richardson-Burns, J. Hendricks, and J. Yang. The morphology of poly(3,4-ethylenedioxythiophene). *Polym. Rev.*, 50:340–384, 2010.
- [411] T. Takano, H. Masunaga, A. Fujiwara, H. Okuzaki, and T. Sasaki. PEDOT nanocrystal in highly conductive PEDOT:PSS polymer films. *Macromolecules*, 45(9):3859–3865, 2012.
- [412] A.M. Nardes, M. Kemerink, R.A.J. Janssen, J.A.M. Bastiaansen, N.M.M. Kiggen, B.M.W. Langeveld, A.J.J.M. van Breemen, and M.M. de Kok. Microscopic understanding of the anisotropic conductivity of PEDOT:PSS thin films. *Adv. Mater.*, 19:1196–2000, 2007.
- [413] X. Crispin, S. Marciniak, W. Osikowicz, G. Zotti, A.W.D. Van der Gon, F. Louwet, M. Fahlman, L. Groenendaal, F. De Schryver, and W.R. Salaneck. Conductivity, morphology, interfacial chemistry, and stability of poly(3,4-ethylene dioxythiophene)-poly(styrene sulfonate): A photoelectron spectroscopy study. *J. Polym. Sci. B*, 41:2561–2583, 2003.
- [414] M. Kemerink, S. Timpanaro, M.M. de Kok, E.A. Meulenkaamp, and F.J. Touwslager. Three-dimensional inhomogeneities in PEDOT:PSS films. *J. Phys. Chem. B*, 108:18820–18825, 2004.
- [415] K. van de Ruit, R.I. Cohen, D. Bollen, T. van Mol, R. Yerushalmi-Rozen, R.A.J. Janssen, and M. Kemerink. Quasi-one dimensional in-plane conductivity in filamentary films of PEDOT:PSS. *Adv. Funct. Mater.*, 23:5778–5786, 2013.
- [416] X. Crispin, F.L.E. Jakobsson, A. Crispin, P.C.M. Grim, P. Andersson, A. Volodin, C. van Haesendonck, M. Van der Auweraer, W.R. Salaneck, and M. Berggren. The origin of the high conductivity of poly(3,4-ethylenedioxythiophene) - poly(styrenesulfonate) (PEDOT-PSS) plastic electrodes. *Chem. Mater.*, 18:4354–4360, 2006.
- [417] L.M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L.P.E. Yunker, J.E. Hein, and A. Aspuru-Guzik. ChemOS: Orchestrating autonomous experimentation. *Sci. Robot.*, 3(19):eaat5559, 2018.
- [418] T. Dimitrov, C. Kreisbeck, J.S. Becker, A. Aspuru-Guzik, and S.K. Saikin. Autonomous molecular design: Then and now. *ACS Appl. Mater. Interfaces*, 11(28):10.1021/acsami.9b01226, 2019.

-
- [419] D. Gelbwaser-Klimovsky, S.K. Saikin, R.H. Goldsmith, and A. Aspuru-Guzik. Optical spectra of p-doped PEDOT nanoaggregates provide insight into the material disorder. *ACS Energy Lett.*, 1:1100–1105, 2016.
- [420] L. Martínez, R. Andrade, E.G. Birgin, and J.M. Martínez. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comb. Chem.*, 30(13):2157–2164, 2009.
- [421] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT 2009.
- [422] L.M. Roch and K.K. Baldrige. Dispersion-corrected spin-component-scaled double-hybrid density functional theory: Implementation and performance for non-covalent interactions. *J. Chem. Theory Comput.*, 13:2650–2666, 2017.
- [423] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case. Development and testing of a general amber force field. *J. Comb. Chem.*, 25(9):1157–1174, 2004.
- [424] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Chem. Phys.*, 117(1):1–19, 1995.
- [425] E.G. Kim and J.L. Brédas. Electronic evolution of poly(3,4-ethylenedioxythiophene) (PEDOT): From the isolated chain to the pristine and heavily doped crystals. *J. Am. Chem. Soc.*, 130(50):16880–16889, 2008.
- [426] H. Li and J.L. Brédas. First-principles theoretical investigation of the electronic couplings in single crystals of phenanthroline-based organic semiconductors. *J. Chem. Phys.*, 126:164704, 2007.
- [427] J.L. Brédas, R. Silbey, D.S. Boudreaux, and R.R. Chance. Chain-length dependence of electronic and electrochemical properties of conjugated systems: polyacetylene, polyphenylene, polythiophene, and polypyrrole. *J. Am. Chem. Soc.*, 105(22):6555–6559, 1983.
- [428] A. Dkhissi, F. Louwet, L. Groenendaal, D. Beljonne, R. Lazzaroni, and J.L. Brédas. Theoretical investigation of the nature of the ground state in the low-bandgap conjugated polymer, poly(3,4-ethylenedioxythiophene). *Chem. Phys. Lett.*, 359(5-6):466–472, 2002.

- [429] R.E. Martin, U. Gubler, J. Cornil, M. Balakina, C. Boudon, C. Bosshard, J.P. Gisselbrecht, F. Diederich, P. Günter, M. Gross, and J.L. Brédas. Monodisperse poly(triacetylene) oligomers extending from monomer to hexadecamer: Joint experimental and theoretical investigation of physical properties. *Chem. Eur. J.*, 6(19):3622–3635, 2000.
- [430] J.J. Apperloo, L. Groenendaal, H. Verheyen, M. Jayakannan, R.A.J. Janssen, A. Dkhissi, D. Beljonne, R. Lazzaroni, and J.L. Brédas. Optical and redox properties of a series of 3,4-ethylenedioxythiophene oligomers. *Chem. Eur. J.*, 8(10):2384–2396, 2002.
- [431] D. Beljonne, J. Cornil, H. Sirringhaus, P.J. Brown, M. Shkunov, R.H. Friend, and J.L. Brédas. Optical signature of delocalized polarons in conjugated polymers. *Adv. Funct. Mater.*, 11(3):229–234, 2001.
- [432] V. Stehr, J. Pfister, R.F. Fink, B. Engels, and C. Deibel. First-principles calculations of anisotropic charge-carrier mobilities in organic semiconductor crystals. *Phys. Rev. B*, 83:155208, 2011.
- [433] A. Kubas, F. Hoffmann, A. Heck, H. Oberhofer, M. Elstner, and J. Blumberger. Electronic couplings for molecular charge transfer: Benchmarking CDFT, FODFT, and FODFTB against high-level ab initio calculations. *J. Chem. Phys.*, 140:104105, 2014.
- [434] R.J. Cave and M.D. Newton. Calculation of electronic coupling matrix elements for ground and excited state electron transfer reactions: Comparison of the generalized Mulliken–Hush and block diagonalization methods. *J. Chem. Phys.*, 106:9213, 1997.
- [435] J.L. Brédas and G.B. Street. Polarons, bipolarons, and solitons in conducting polymers. *Acc. Chem. Res.*, 1305(4):309–315, 1985.
- [436] J.D. Wright. *Molecular Crystals*. Cambridge, UK: Cambridge University Press, 1995.
- [437] F. Bassani and G.P. Parravicini. *Electronic states and optical transitions in solids*. Oxford: Pergamon, 1975.
- [438] J.L. Brédas, D. Beljonne, V. Coropceanu, and J. Cornil. Charge-transfer and energy-transfer processes in π -conjugated oligomers and polymers: A molecular picture. *Chem. Rev.*, 104:4971–5003, 2004.
- [439] E.H. Horak, M.T. Rea, K.D. Heylman, D. Gelbwaser-Klimovsky, S.K. Saikin, B.J. Thompson, D.D. Kohler, K.A. Knapper, W. Wei, F. Pan, P. Gopalan, J.C. Wright, A. Aspuru-Guzik, and R.H. Goldsmith. Exploring electronic structure and order in polymers via single-particle microresonator spectroscopy. *Nano Lett.*, 18:1600–1607, 2018.
- [440] J.R. O’Dea, L.M. Brown, N. Hoepker, J.A. Marohn, and S. Sadewasser. Scanned probe microscopy of solar cells: From inorganic thin films to organic photovoltaics. *MRS Bull.*, 37:642–650, 2012.

- [441] C. Groves, O.G. Reid, and D.S. Ginger. Heterogeneity in polymer solar cells: Local morphology and performance in organic photovoltaics studied with scanning probe microscopy. *Acc. Chem. Res.*, 43:612–620, 2010.
- [442] X. Gao and A. Eisfeld. Near-field spectroscopy of nanoscale molecular aggregates. *J. Phys. Chem. Lett.*, 9:6003–6010, 2018.
- [443] S.B. Penwell, L.D.S. Ginsberg, R. Noriega, and N.S. Ginsberg. Resolving ultrafast exciton migration in organic solids at the nanoscale. *Nat. Mater.*, 16:1136–1141, 2017.
- [444] P.F. Barbara, A.J. Gesquiere, S.J. Park, and Y.J. Lee. Single-molecule spectroscopy of conjugated polymers. *Acc. Chem. Res.*, 38:602–610, 2005.
- [445] A. Thiessen, J. Vogelsang, T. Adachi, F. Steiner, D. Vanden Bout, and J. Lupton. Unraveling the chromophoric disorder of poly(3-hexylthiophene). *Proc. Natl. Acad. Sci.*, 110:E3550–E3556, 2013.
- [446] K.D. Heylman, N. Thakkar, E.H. Horak, S.C. Quillin, C. Cherqui, K.A. Knapper, D.J. Masiello, and R.H. Goldsmith. Optical microresonators as single-particle absorption spectrometers. *Nat. Photon.*, 10:788–795, 2016.
- [447] M.T. Rea, F. Pan, E. Horak, K.A. Knapper, H.L. Nguyen, C.H. Vollbrecht, and R.H. Goldsmith. Investigating the mechanism of post-treatment on PEDOT:PSS via single-particle absorption spectroscopy. *J. Phys. Chem. C*, 2019.
- [448] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.
- [449] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Internat. Conf. Mach. Learn.*, volume 96, pages 148–156. Citeseer, 1996.
- [450] C.M. Bishop. Mixture density networks. Technical report, Citeseer, 1994.
- [451] F. Richter. Mixture density networks. In *Kombination Künstlicher Neuronaler Netze*, pages 171–198. Springer, 2003.
- [452] A. Aspuru-Guzik, R. Lindh, and M. Reiher. The matter simulation (r)evolution. *ACS Cent. Sci.*, 4(2):144–152, 2018.
- [453] I. Navizet, Y.J. Liu, N. Ferré, D. Roca-Sanjuán, and R. Lindh. The chemistry of bioluminescence: An analysis of chemical functionalities. *ChemPhysChem*, 12(17):3064–3076, 2011.
- [454] M. Vacher, I. Fdez. Galván, B.W. Ding, S. Schramm, R. Berraud-Pache, P. Naumov, N. Ferré, Y.J. Liu, I. Navizet, D. Roca-Sanjuán, W.J. Baader, and R. Lindh. Chemi- and bioluminescence of cyclic peroxides. *Chem. Rev.*, 118(15):6927–6974, 2018.
- [455] C. Dodeigne, L. Thunus, and R. Lejeune. Chemiluminescence as diagnostic tool. A review. *Talanta*, 51(3):415–439, 2000.

- [456] M. Ronaghi, M. Uhlén, and P. Nyrén. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363–365, 1998.
- [457] A. Mayer and S. Neuenhofer. Luminescent labels—more than just an alternative to radioisotopes? *Angew. Chem. Int. Ed.*, 33(10):1044–1072, 1994.
- [458] Y. Chen, A.J.H. Spiering, S. Karthikeyan, G.W.M. Peters, E.W. Meijer, and R.P. Sijbesma. Mechanically induced chemiluminescence from polymers incorporating a 1, 2-dioxetane unit in the main chain. *Nat. Chem.*, 4(7):559–562, 2012.
- [459] J.M. Clough, A. Balan, T.L.J. van Daal, and R.P. Sijbesma. Probing force with mechanobase-induced chemiluminescence. *Angew. Chem. Int. Ed.*, 55(4):1445–1449, 2016.
- [460] L. De Vico, Y.J. Liu, J.W. Wisborg Krogh, and R. Lindh. Chemiluminescence of 1,2-dioxetane. reaction mechanism uncovered. 111(32):8013–8019, 2007.
- [461] P. Farahani, D. Roca-Sanjuán, F. Zapata, and R. Lindh. Revisiting the nonadiabatic process in 1,2-dioxetane. *J. Chem. Theory Comput.*, 9(12):5404–5411, 2013.
- [462] M. Vacher, A. Brakestad, H.O. Karlsson, I. Fdez. Galván, and R. Lindh. Dynamical insights into the decomposition of 1, 2-dioxetane. *J. Chem. Theory Comput.*, 13(6):2448–2457, 2017.
- [463] M. Vacher, P. Farahani, A. Valentini, L.M. Frutos, H.O. Karlsson, I. Fdez. Galvaán, and R. Lindh. How do methyl groups enhance the triplet chemiexcitation yield of dioxetane? *J. Phys. Chem. Lett.*, 8(16):3790–3794, 2017.
- [464] V. Botu and R. Ramprasad. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.*, 115(16):1074–1083, 2015.
- [465] Z. Li, J.R. Kermode, and A. de Vita. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.*, 114(9):096405, 2015.
- [466] F. Brockherde, L. Vogt, L. Li, M.E. Tuckerman, K. Burke, and K.R. Müller. Bypassing the kohn-sham equations with machine learning. *Nat. Commun.*, 8(1):872, 2017.
- [467] E. Schneider, L. Dai, R.Q. Topper, C. Drechsel-Grau, and M.E. Tuckerman. Stochastic neural network approach for learning high-dimensional free energy surfaces. *Phys. Rev. Lett.*, 119(15):150601, 2017.
- [468] P.O. Dral, M. Barbatti, and W. Thiel. Nonadiabatic excited-state dynamics with machine learning. *J. Phys. Chem. Lett.*, 9(19):5660–5663, 2018.
- [469] W.K. Chen, X.Y. Liu, W.H. Fang, P.O. Dral, and G. Cui. Deep learning for nonadiabatic excited-state dynamics. *J. Phys. Chem. Lett.*, 9(23):6702–6708, 2018.
- [470] R.M. Balabin and E.I. Lomakina. Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. *J. Chem. Phys.*, 131(7):074104, 2009.

- [471] R. Ramakrishnan, P.O. Dral, M. Rupp, and O.A. von Lilienfeld. Big data meets quantum chemistry approximations: The Δ -machine learning approach. *J. Chem. Theory Comput.*, 11(5):2087–2096, 2015.
- [472] M. Rupp, A. Tkatchenko, K.R. Müller, and O.A. von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108(5):058301, 2012.
- [473] M. Gastegger, J. Behler, and P. Marquetand. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.*, 8(10):6924–6935, 2017.
- [474] B.O. Roos, P.R. Taylor, and P.E.M. Sigbahn. A complete active space SCF method (CASSCF) using a density matrix formulated super-CI approach. 48(2):157–173, 1980.
- [475] Björn O Roos. The complete active space self-consistent field method and its applications in electronic structure calculations. *Advances in Chemical Physics: Ab Initio Methods in Quantum Chemistry Part 2*, 69:399–445, 1987.
- [476] B.O. Roos, R. Lindh, P.A. Malmqvist, V. Veryazov, and P.O. Widmark. Main group atoms and dimers studied with a new relativistic ANO basis set. 108(15):2851–2858, 2004.
- [477] F. Aquilante, L. Gagliardi, T.B. Pedersen, and R. Lindh. Atomic cholesky decompositions: A route to unbiased auxiliary basis sets for density fitting approximation with tunable accuracy and efficiency. *J. Chem. Phys.*, 130(15):154107, 2009.
- [478] F. Aquilante, J. Autschbach, R.K. Carlson, L.F. Chibotaru, M.G. Delcey, L. De Vico, I. Fdez. Galván, N. Ferré, L.M. Frutos, L. Gagliardi, A.G. Garavelli, C.E. Hoyer, G. Li Manni, H. Lischka, D. Ma, P.A. Malmqvist, T. Müller, A. Nenov, M. Olivucci, T.B. Pedersen, D. Peng, F. Plasser, B. Pritchard, M. Reiher, I. Rivalta, I. Schapiro, J. Segarra-Martí, M. Stenrup, D.G. Truhlar, L. Ungur, A. Valentini, S. Vancoillie, V. Veryazov, V.P. Vysotskiy, O. Weingart, F. Zapata, and R. Lindh. Molcas 8: New capabilities for multiconfigurational quantum chemical calculations across the periodic table. *J. Comp. Chem.*, 37(5):506–541, 2016.
- [479] U. Lourderaj, K. Park, and W.L. Hase. Classical trajectory simulations of post-transition state dynamics. *Int. Rev. Phys. Chem.*, 27(3):361–403, 2008.
- [480] C.K.I. Williams. Computing with infinite networks. In *Adv. Neural. Inf. Process. Syst.*, pages 295–301, 1997.
- [481] K.F.R.S. Pearson. Principal components analysis. *Lond. Edinb. Dubl. Phil. Mag.*, 6(2):559, 1901.
- [482] D. Tran, A. Kucukelbir, A.B. Dieng, M. Rudolph, D. Liang, and D.M. Blei. Edward: A Library for Probabilistic Modeling, Inference, and Criticism. *arXiv preprint arXiv:1610.09787*, 2016.
- [483] M. Rupp. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.*, 115(16):1058–1073, 2015.

- [484] M. Rupp, R. Ramakrishnan, and O.A. von Lilienfeld. Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.*, 6(16):3309–3313, 2015.
- [485] K. Vu, J.C. Snyder, L. Li, M. Rupp, B.F. Chen, T. Khelif, K.R. Müller, and K. Burke. Understanding kernel ridge regression: Common behaviors from simple functions to density functionals. *Int. J. Quantum Chem.*, 115(16):1115–1128, 2015.
- [486] P.M. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Comput.*, 7(1):117–143, 1995.
- [487] S. Mohamed, K. Heller, and Z. Ghahramani. Bayesian and L1 approaches to sparse unsupervised learning. In *Internat. Conf. Mach. Learn.*, 2012.
- [488] R. Abegg. Die valenz und das periodische system. versuch einer theorie der molekularverbindungen. *Z. Anorg. Allg. Chem.*, 39(1):330–380, 1904.
- [489] G.N. Lewis. The atom and the molecule. *J. Am. Chem. Soc.*, 38(4):762–785, 1916.
- [490] I. Langmuir. The arrangement of electrons in atoms and molecules. *J. Am. Chem. Soc.*, 41(6):868–934, 1919.
- [491] L. Pauling, L.O. Brockway, and J.Y. Beach. The dependence of interatomic distance on single bond-double bond resonance. *J. Am. Chem. Soc.*, 57(12):2705–2709, 1935.
- [492] C.A. Coulson. The electronic structure of some polyenes and aromatic molecules. vii. bonds of fractional order by the molecular orbital method. *Proc. R. Soc. London*, 169(938):413–428, 1939.
- [493] L. Pauling. The nature of the chemical bond. Application of results obtained from the quantum mechanics and from a theory of paramagnetic susceptibility to the structure of molecules. *J. Am. Chem. Soc.*, 53(4):1367–1400, 1931.
- [494] R.J. Gillespie and R.S. Nyholm. Inorganic stereochemistry. *Quarterly Reviews, Chemical Society*, 11(4):339–380, 1957.
- [495] R.J. Gillespie. The electron-pair repulsion model for molecular geometry. *J. Chem. Ed.*, 47(1):18, 1970.
- [496] W. Adam and W.J. Baader. Effects of methylation on the thermal stability and chemiluminescence properties of 1,2-dioxetanes. *J. Am. Chem. Soc.*, 107(2):410–416, 1985.
- [497] S.A. Lopez, B. Sanchez-Lengeling, J. de Goes Soares, and A. Aspuru-Guzik. Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule*, 1(4):857–870, 2017.
- [498] K.F. Jensen, B.J. Reizman, and S.G. Newman. Tools for chemical synthesis in microsystems. *Lab Chip*, 14(17):3206–3212, 2014.

- [499] J.N. Wei, D. Duvenaud, and A. Aspuru-Guzik. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.*, 2(10):725–732, 2016.
- [500] C.W. Coley, R. Barzilay, T.S. Jaakkola, W.H. Green, and K.F. Jensen. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.*, 3:434–443, 2017.
- [501] M.H.S. Segler, T. Kogej, C. Tyrchan, and M.P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.*, 4(1):120–131, 2017.
- [502] P. Nikolaev, D. Hooper, N. Perea-Lopez, M. Terrones, and B. Maruyama. Discovery of wall-selective carbon nanotube growth conditions via automated experimentation. *ACS Nano*, 8(10):10214–10222, 2014.
- [503] D.E. Fitzpatrick, C. Battilocchio, and S.V. Ley. A novel internet-based reaction monitoring, control and autonomous self-optimization platform for chemical synthesis. *Org. Proc. Res. Dev.*, 20(2):386–394, 2016.
- [504] V. Duros, J. Grizou, W. Xuan, Z. Hosni, D.L. Long, H.N. Miras, and L. Cronin. Human versus robots in the discovery and crystallization of gigantic polyoxometalates. *Angew. Chem. Int. Ed.*, 129(36):10955–10960, 2017.
- [505] J.A. Bergstra, R. Badenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Adv. Neural. Inf. Process. Syst.*, pages 2546–2554, 2011.
- [506] T. Back. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford university press, New York, 1996.
- [507] D. Simon. *Evolutionary Optimization Algorithms*. John Wiley & Sons, New York, 2013.
- [508] Z. Zhou, X. Li, and R.N. Zare. Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci.*, 3(12):1337–1344, 2017.
- [509] J. Snoek, K. Swersky, R. Zemel, and R.P. Adams. Input warping for Bayesian optimization of non-stationary functions. In *Internat. Conf. Mach. Learn.*, pages 1674–1682, 2014.
- [510] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.*, 4(2):268–276, 2018.
- [511] E.O. Pyzer-Knapp, G.N. Simm, and A. Aspuru-Guzik. A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. *Mater. Horizons*, 3(3):226–233, 2016.
- [512] S. Ju, T. Shiga, L. Feng, Z. Hou, K. Tsuda, and J. Shiomi. Designing nanostructures for phonon transport via Bayesian optimization. *Phys. Rev. X*, 7(2):021024, 2017.

- [513] Z. Wang, C. Li, S. Jegelka, and P. Kohli. Batched high-dimensional Bayesian optimization via structural kernel learning. In *Internat. Conf. Mach. Learn.*, pages 3656–3664, 2017.
- [514] Z. Wang, C. Gehring, P. Kohli, and S. Jegelka. Batched large-scale Bayesian optimization in high-dimensional spaces. In *Internat. Conf. Artif. Intell. Stat.*, pages 745–754, 2018.
- [515] S. Marmin, C. Chevalier, and D. Ginsbourger. Efficient batch-sequential Bayesian optimization with moments of truncated Gaussian vectors. *arXiv preprint arXiv:1609.02700*, 2016.
- [516] J. González, M. Osborne, and N. Lawrence. GLASSES: Relieving the myopia of Bayesian optimisation. In *Artif. Intell. Stat.*, pages 790–799, 2016.
- [517] M.D. Hoffman and A. Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15:1593–1623, 2014.
- [518] J. Salvatier, T.V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Comp. Sci.*, 2:e55, 2016.
- [519] E. A. Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.
- [520] G.S. Watson. Smooth regression analysis. *Sankhya A.*, pages 359–372, 1964.
- [521] L.J.V. Miranda. PySwarms, a research-toolkit for Particle Swarm Optimization in Python. *J. of Open Source Soft.*, 3, 2018.
- [522] N. Hansen. A Python implementation of CMA-ES. <https://github.com/CMA-ES/pycma>, 2018.
- [523] R.J. Field and R.M. Noyes. Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction. *J. Chem. Phys.*, 60(5):1877–1884, 1974.
- [524] N.G. van Kampen. *Stochastic Processes in Physics and Chemistry*, volume 1. Elsevier, Amsterdam, 1992.
- [525] A.M. Zhabotinsky. A history of chemical oscillations and waves. *Chaos*, 1:379–386, 1991.
- [526] H. Deng. Effect of bromine derivatives of malonic acid on the oscillating reaction of malonic acid, cerium ions and bromate. *Nature*, 213:589–590, 1967.
- [527] A.M. Zhabotinsky and A.N. Zaikin. Oscillatory processes in biological and chemical systems. *Izdatelstro "Nauka" Publishers, Moscow*, 1967.
- [528] L. Gyorgyi, R. Turányi, and R. J. Field. Mechanistic details of the oscillatory Belousov-Zhabotinski reaction. *J. Phys. Chem.*, 94:7162–7170, 1990.
- [529] R.J. Field. Das Experiment: Eine oszillierende Reaktion. *Chem. unserer Zeit*, 7:171–176, 1973.

- [530] K. Bar-Eli. The minimal bromate oscillator simplified. *J. Phys. Chem.*, 89:2855–2860, 1985.
- [531] V. Voorsluijs, I. G. Kevrekidis, and Y. De Decker. Nonlinear behavior and fluctuation-induced dynamics in the photosensitive Belousov–Zhabotinsky reaction. *Phys. Chem. Chem. Phys.*, 19:22528–22537, 2017.
- [532] A. Peruzzo, J. McClean, P. Shadbolt, M.H. Yung, X.Q. Zhou, P.J. Love, A. Aspuru-Guzik, and J.L. O’Brien. A variational eigenvalue solver on a photonic quantum processor. *Nat. Commun.*, 5:4213, 2014.
- [533] J. Romero, J.P. Olson, and A. Aspuru-Guzik. Quantum autoencoders for efficient compression of quantum data. *Quant. Sci. Technol.*, 2(4):045001, 2017.
- [534] J.P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V.R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, and T. Buonassisi. Accelerating materials development via automation, machine learning, and high-performance computing. *Joule*, 2(8):1410–1420, 2018.
- [535] H. Koinuma and I. Takeuchi. Combinatorial solid-state chemistry of inorganic materials. *Nat. Mater.*, 3(7):429, 2004.
- [536] S.M. Mennen, C. Alhambra, C.L. Allen, M. Barberis, S. Berritt, T.A. Brandt, A.D. Campbell, J. Castañón, A.H. Cherney, M. Christensen, D.B. Damon, J.E. de Diego, S. García-Cerrada, P. Garía-Losada, R. Haro, J. Janey, D.C. Leitch, L. Li, F. Lui, P.C. Lobben, D.W.C. MacMillan, J. Magano, E. McInturff, S. Monfette, R.J. Post, D. Schultz, B.J. Sitter, J.M. Stevens, I.I. Strambeanu, J. Twilton, K. Wang, and M.A. Zajac. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Org. Proc. Res. Dev.*, 23(6):1213–1242, 2019.
- [537] K. Troshin and J.F. Hartwig. Snap deconvolution: An informatics approach to high-throughput discovery of catalytic reactions. *Science*, 357(6347):175–181, 2017.
- [538] J.A. Selekman, J. Qiu, K. Tran, J. Stevens, V. Rosso, E. Simmons, Y. Xiao, and J. Janey. High-throughput automation in chemical process development. *Annu. Rev. Chem. Biomol. Eng.*, 8:525–547, 2017.
- [539] K.D. Collins, T. Gensch, and F. Glorius. Contemporary screening approaches to reaction discovery and development. *Nat. Chem.*, 6(10):859, 2014.
- [540] A. McNally, C.K. Prier, and D.W.C. MacMillan. Discovery of an α -amino C–H arylation reaction using the strategy of accelerated serendipity. *Science*, 334(6059):1114–1117, 2011.
- [541] K.W. Moore, A. Pechen, X.J. Feng, J. Dominy, V.J. Beltrani, and H. Rabitz. Why is chemical synthesis and property optimization easier than expected? *Phys. Chem. Chem. Phys.*, 13(21):10048–10070, 2011.

- [542] M. Yoshida, T. Hinkley, S. Tsuda, Y.M. Abul-Haija, R.T. McBurney, V. Kulikov, J.S. Mathieson, S.G. Reyes, M.D. Castro, and L. Cronin. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem*, 4(3):533–543, 2018.
- [543] C.W. Coley, D.A. Thomas, J.A.M. Lummiss, J.N. Jaworski, C.P. Breen, V. Schultz, T. Hart, J.S. Fishman, L. Rogers, H. Gao, R.W. Hicklin, P.P. Plehiers, J. Byington, J.S. Piotti, W.H. Gren, A.J. Hart, T.F. Jamison, and K.F. Jensen. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*, 365(6453):eaax1566, 2019.
- [544] F. Ren, L. Ward, T. Williams, K.J. Laws, C. Wolverton, J. Hattrick-Simpers, and A. Mehta. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.*, 4(4):eaq1566, 2018.
- [545] M.J. Casciato, S. Kim, J.C. Lu, D.W. Hess, and M.A. Grover. Optimization of a carbon dioxide-assisted nanoparticle deposition process using sequential experimental design with adaptive design space. *Ind. Eng. Chem.*, 51(11):4363–4370, 2012.
- [546] P.B. Wigley, P.J. Everitt, A. van den Hengel, J.W. Bastian, M.A. Sooriyabandara, G.D. McDonald, K.S. Hardman, C.D. Quinlivan, P. Manju, C.C.N. Kuhn, I.R. Petersen, A.N. Luiten, J.J. Hope, N.P. Robins, and M.R. Hush. Fast machine-learning online optimization of ultra-cold-atom experiments. *Sci. Rep.*, 6:25890, 2016.
- [547] A.C. Bédard, A. Adamo, K.C. Aroh, M.G. Russell, A.A. Bedermann, J. Torosian, B. Yue, K.F. Jensen, and T.F. Jamison. Reconfigurable system for automated optimization of diverse chemical reactions. *Science*, 361(6408):1220–1225, 2018.
- [548] D. Cortés-Borda, E. Wimmer, B. Gouilleux, E. Barré, N. Oger, L. Goulamaly, L. Peault, B. Charrier, C. Truchet, P. Giraudeau, M. Rodrigues-Zubiri, E. Le Grogneq, and F.X. Felpin. An autonomous self-optimizing flow reactor for the synthesis of natural product carpanone. *J. Org. Chem.*, 83(23):14286–14299, 2018.
- [549] B. Maruyama, K. Decker, P. Nikolaev, M. Krein, J. Poleski, and R. Barto. Autonomous experimentation applied to carbon nanotube synthesis. In *Meeting Abstracts*, number 9, pages 668–668. The Electrochemical Society, 2017.
- [550] M.M. Noack, K.G. Yager, M. Fukuto, G.S. Doerk, R. Li, and J.A. Sethian. A kriging-based approach to autonomous experimentation with applications to X-ray scattering. *Sci. Rep.*, 9(1):1–19, 2019.
- [551] A.G. Kusne, T. Gao, A. Mehta, L. Ke, M.C. Nguyen, K.M. Ho, V. Antropov, C.Z. Wang, M.J. Kramer, C. Long, and I. Takeuchi. On-the-fly machine-learning for high-throughput experiments: Search for rare-earth-free permanent magnets. *Sci. Rep.*, 4:6367, 2014.
- [552] J. Li, Y. Tu, R. Liu, Y. Lu, and X. Zhu. Toward “on-demand” materials synthesis and scientific discovery through intelligent robots. *Adv. Sci.*, page 1901957, 2020.

-
- [553] L.M. Baumgartner, C.W. Coley, B.J. Reizman, K.W. Gao, and K.F. Jensen. Optimum catalyst selection over continuous and discrete process variables with a single droplet microfluidic reaction platform. *React. Chem. Eng.*, 3(3):301–311, 2018.
- [554] B.J. Reizman and K.F. Jensen. Simultaneous solvent screening and reaction optimization in microliter slugs. *Chem. Comm.*, 51(68):13290–13293, 2015.
- [555] D.J. Lizotte, T. Wang, M.H. Bowling, and D. Schuurmans. Automatic gait optimization with Gaussian process regression. In *IJCAI*, volume 7, pages 944–949, 2007.
- [556] Y. Bai, L. Wilbraham, B.J. Slater, M.A. Zwijnenburg, R.S. Sprick, and A.I. Cooper. Accelerated discovery of organic polymer photocatalysts for hydrogen evolution from water through the integration of experiment and theory. *J. Am. Chem. Soc.*, 141(22):9063–9071, 2019.
- [557] D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1487–1495. ACM, 2017.
- [558] The GPyOpt Authors. GPyOpt: A Bayesian optimization framework in Python. <http://github.com/SheffieldML/GPyOpt>, 2016.
- [559] E.C. Garrido-Merchán and D. Hernández-Lobato. Dealing with categorical and integer-valued variables in Bayesian optimization with Gaussian processes. *Neurocomp.*, 2019.
- [560] J.M. Hernández-Lobato, J. Requeima, E.O. Pyzer-Knapp, and A. Aspuru-Guzik. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. In *Internat. Conf. Mach. Learn.*, pages 1470–1479. JMLR. org, 2017.
- [561] K.W. Ng, G.L. Tian, and M.L. Tang. *Dirichlet and related distributions: Theory, methods and applications*, volume 888. John Wiley & Sons, 2011.
- [562] J. Atchison and S.M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.
- [563] D.P. Kingma and M. Welling. Auto-encoding variational Bayes. 2014.
- [564] C.J. Maddison, A. Mnih, and Y.W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.
- [565] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with Gumbel-softmax. 2017.
- [566] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Comput.*, 4(1):1–58, 1992.
- [567] C. Molnar. *Interpretable machine learning*. Lulu.com, 2019.
- [568] C.S. Perone. PyEvolve: a Python open-source framework for genetic algorithms. *ACM Sigevolution*, 4(1):12–20, 2009.

- [569] J. González, Z. Dai, P. Hennig, and N. Lawrence. Batch Bayesian optimization via local penalization. In *Artif. Intell. Stat.*, pages 648–657, 2016.
- [570] J. Gonzalez, J. Longworth, D.C. James, and N.D. Lawrence. Bayesian optimization for synthetic gene design. *arXiv preprint arXiv:1505.01627*, 2015.
- [571] J.S. Bergstra, D. Yamins, and D.D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. 2013.
- [572] R.E. Bellman. *Adaptive control processes: a guided tour*, volume 2045. Princeton university press, 2015.
- [573] H. Sahu, W. Rao, A. Troisi, and H. Ma. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Adv. Energy Mater.*, 8(24):1801032, 2018.
- [574] V. Venkatraman and B.K. Alsberg. A quantitative structure-property relationship study of the photovoltaic performance of phenothiazine dyes. *Dyes Pigm.*, 114:69–77, 2015.
- [575] M.C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A.J. Heeger, and C.J. Brabec. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Adv. Mater.*, 18(6):789–794, 2006.
- [576] T. Ameri, G. Dennler, C. Lungenschmied, and C.J. Brabec. Organic tandem solar cells: A review. *Energy Environ. Sci.*, 2(4):347–363, 2009.
- [577] C. Kim, T.D. Huan, S. Krishnan, and R. Ramprasad. A hybrid organic-inorganic perovskite dataset. *Sci. Data*, 4:170057, 2017.
- [578] J. Kim, S.H. Lee, C.H. Chung, and K.H. Hong. Systematic analysis of the unique band gap modulation of mixed halide perovskites. *Phys. Chem. Chem. Phys.*, 18(6):4423–4428, 2016.
- [579] J. Kim, S.C. Lee, S.H. Lee, and K.H. Hong. Importance of orbital interactions in determining electronic band structures of organo-lead iodide. *J. Phys. Chem. C*, 119(9):4627–4634, 2015.
- [580] P. Umari, E. Mosconi, and F. De Angelis. Relativistic GW calculations on $\text{CH}_3\text{NH}_3\text{PbI}_3$ and $\text{CH}_3\text{NH}_3\text{SnI}_3$ perovskites for solar cell applications. *Sci. Rep.*, 4:4467, 2014.
- [581] G. Giorgi, J.I. Fujisawa, H. Segawa, and K. Yamashita. Cation role in structural and electronic properties of 3D organic–inorganic halide perovskites: a DFT analysis. *J. Phys. Chem. C*, 118(23):12176–12183, 2014.
- [582] J. Endres, D.A. Egger, M. Kulbak, R.A. Kerner, L. Zhao, S.H. Silver, G. Hodes, B.P. Rand, D. Cahen, L. Kronik, and A. Kahn. Valence and conduction band densities of states of metal halide perovskites: a combined experimental–theoretical study. *J. Phys. Chem. Lett.*, 7(14):2722–2729, 2016.

-
- [583] N. Miyaura. Metal-catalyzed cross-coupling reactions of organoboron compounds with organic halides. *Metal-Catalyzed Cross-Coupling Reactions*, pages 41–123, 2004.
- [584] D.G. Brown and J. Bostrom. Analysis of past and present synthetic methodologies on medicinal chemistry: Where have all the new reactions gone? miniperspective. *J. Med. Chem.*, 59(10):4443–4458, 2015.
- [585] M. Awad and R. Khanna. *Multiobjective Optimization*, pages 185–208. Apress, Berkeley, CA, 2015.
- [586] L.M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L.P.E. Yunker, J.E. Hein, and A. Aspuru-Guzik. ChemOS: An orchestration software to democratize autonomous discovery. *PLoS one*, 15(4):e0229862, 2020.
- [587] D. Caramelli, D. Salley, A. Henson, G.A. Camarasa, S. Sharabi, G. Keenan, and L. Cronin. Networking chemical robots for reaction multitasking. *Nat. Commun.*, 9(1):1–10, 2018.
- [588] V. Sans, L. Porwol, V. Dragone, and L. Cronin. A self optimizing synthetic organic reactor system using real-time in-line NMR spectroscopy. *Chem. Sci.*, 6(2):1258–1264, 2015.
- [589] M. Trobe and M.D. Burke. The molecular industrial revolution: Automated synthesis of small molecules. *Angew. Chem. Int. Ed.*, 57(16):4192–4214, 2018.
- [590] P.R. Wiecha, A. Arbouet, C.N. Girard, A. Lecestre, G. Larrieu, and V. Paillard. Evolutionary multi-objective optimization of colour pixels based on dielectric nanoantennas. *Nat. Nanotechnol.*, 12(2):163, 2017.
- [591] J. Jung. Robust design of plasmonic waveguide using gradient index and multiobjective optimization. *IEEE Photonics Technol. Lett.*, 28(7):756–758, 2016.
- [592] M.H. Ahmadi, M.A. Ahmadi, R. Bayat, M. Ashouri, and M. Feidt. Thermo-economic optimization of stirling heat pump by using non-dominated sorting genetic algorithm. *Energy Convers. Manag.*, 91:315–322, 2015.
- [593] O. Maaliou and B.J. McCoy. Optimization of thermal energy storage in packed columns. *Sol. Energy*, 34(1):35–41, 1985.
- [594] A.J. White, J.D. McTigue, and C.N. Markides. Analysis and optimisation of packed-bed thermal reservoirs for electricity storage applications. *Proc. Inst. Mech. Eng. A*, 230(7):739–754, 2016.
- [595] J. Marti, L. Geissbühler, V. Becattini, A. Haselbacher, and A. Steinfeld. Constrained multi-objective optimization of thermocline packed-bed thermal-energy storage. *Appl. Energy*, 2018.
- [596] H.M. Dubey, M. Pandit, and B.K. Panigrahi. Hydro-thermal-wind scheduling employing novel ant lion optimization technique with composite ranking index. *Renew. Energy*, 99:18–34, 2016.

- [597] D.N. Jumbam, R.A. Skilton, A.J. Parrott, R.A. Bourne, and M. Poliakoff. The effect of self-optimisation targets on the methylation of alcohols using dimethyl carbonate in supercritical CO₂. *J. Flow Chem.*, 2(1):24–27, 2012.
- [598] S. Krishnadasan, R.J.C. Brown, A.J. deMello, and J.C. deMello. Intelligent routes to the controlled synthesis of nanoparticles. *Lab Chip*, 7(11):1434–1441, 2007.
- [599] J.S. Moore and K.F. Jensen. Automated multitrajectory method for reaction optimization in a microfluidic system using online IR analysis. *Org. Proc. Res. Dev.*, 16(8):1409–1415, 2012.
- [600] R.T. Marler and J.S. Arora. The weighted sum method for multi-objective optimization: New insights. *Struct. Multidiscipl. Optim.*, 41(6):853–862, 2010.
- [601] T.C. Malig, J.D.B. Koenig, H. Situ, N.K. Chehal, P.G. Hultin, and J.E. Hein. Real-time HPLC-MS reaction progress monitoring using an automated analytical platform. *React. Chem. Eng.*, 2:309, 2017.
- [602] C. Kreisbeck, T. Kramer, and A. Aspuru-Guzik. Disentangling electronic and vibronic coherences in two-dimensional echo spectra. *J. Phys. Chem. B*, 117:9380–9385, 2013.
- [603] C. Kreisbeck and T. Kramer. Exciton dynamics lab for light-harvesting complexes (GPU-HEOM), Feb 2013.
- [604] D. Segal. *Materials for the 21st Century*. Oxford University Press, 2017.
- [605] K. Lin, R. Gómez-Bombarelli, E.S. Beh, L. Tong, Q. Chen, A. Valle, A. Aspuru-Guzik, M.J. Aziz, and R.G. Gordon. A redox-flow battery with an alloxazine-based organic electrolyte. *Nat. Energy*, 1(9):16102, 2016.
- [606] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T.D. Hirzel, D. Duvenaud, D. Maclaurin, M.A. Blood-Forsythe, H.S. Chae, M. Einzinger, D.G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. uang, S.I. Hong, M. Baldo, R.P. Adams, and A. Aspuru-Guzik. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.*, 15(10):1120, 2016.
- [607] C. Porte, W. Debreuille, F. Draskovic, and A. Delacroix. Automation and optimization by simplex methods of 6-chlorohexanol synthesis. *Process Control Qual.*, 4(8):111–122, 1996.
- [608] R.W. Wagner, F. Li, H. Du, and J.S. Lindsey. Investigation of cocatalysis conditions using an automated microscale multireactor workstation: Synthesis of meso-retramesitylporphyrin. *Org. Proc. Res. Dev.*, 3(1):28–37, 1999.
- [609] D.C. Fabry, E. Sugiono, and M. Rueping. Online monitoring and analysis for autonomous continuous flow self-optimizing reactor systems. *React. Chem. Eng.*, 1(2):129–133, 2016.

-
- [610] D.C. Fabry, E. Sugiono, and M. Rueping. Self-optimizing reactor systems: algorithms, on-line analytics, setups, and strategies for accelerating continuous flow process optimization. *Isr. J. Chem.*, 54(4):341–350, 2014.
- [611] D.E. Fitzpatrick, C. Battilocchio, and S.V. Ley. Enabling technologies for the future of chemical synthesis. *ACS Cent. Sci.*, 2(3):131–138, 2016.
- [612] S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, and O. Levy. The high-throughput highway to computational materials design. *Nat. Mater.*, 12(3):191, 2013.
- [613] J. Bajorath. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.*, 1(11):882, 2002.
- [614] H. Altae-Tran, B. Ramsundar, A.S. Pappu, and V. Pande. Low data drug discovery with one-shot learning. *ACS Cent. Sci.*, 3(4):283–293, 2017.
- [615] C.W. Coley, W.H. Green, and K.F. Jensen. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.*, 51(5):1281–1289, 2018.
- [616] M.H.S. Segler, M. Preuss, and M.P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604, 2018.
- [617] W. Jin, C. Coley, R. Barzilay, and T. Jaakkola. Predicting organic reaction outcomes with Weisfeiler-Lehman network. In *Adv. Neural. Inf. Process. Syst.*, pages 2607–2616, 2017.
- [618] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, and V. Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.*, 3(10):1103–1113, 2017.
- [619] D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.*, 7:11241, 2016.
- [620] J.M. Granda, L. Donina, V. Dragone, D.L. Long, and L. Cronin. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*, 559(7714):377, 2018.
- [621] V. Dragone, V. Sans, A.B. Henson, J.M. Granda, and L. Cronin. An autonomous organic reaction search engine for chemical reactivity. *Nat. Commun.*, 8:15733, 2017.
- [622] A.B. Henson, P.S. Gromski, and L. Cronin. Designing algorithms to aid discovery by chemical robots. *ACS Cent. Sci.*, 4(7):793–804, 2018.
- [623] R.A. Skilton, R.A. Bourne, Z. Amara, R. Horvath, J. Jin, M. Scully, E. Streng, S.L.Y. Tang, P. Summers, J. Wang, E. Pérez, N. Asfaw, G.L.P. Aydos, J. Dupont, M.W. Comak, G. and George, and M. Poliakoff. Remote-controlled experiments with cloud chemistry. *Nat. Chem.*, 7:1–5, 2015.

- [624] B.D. Ripley. *Pattern recognition and neural networks*. Cambridge university press, 2007.
- [625] L. Cronin, S.H.M. Mehr, and J.M. Granda. Catalyst: The metaphysics of chemical reactivity. *Chem*, 4(8):1759–1761, 2018.
- [626] R.L. Greenaway, V. Santolini, M.J. Bennison, B.M. Alston, C.J. Pugh, M.A. Little, M. Miklitz, E.G.B. Eden-Rump, R. Clowes, A. Shakil, H.J. Cuthbertson, H. Armstrong, M.E. Briggs, K.E. Jelfs, and A.I. Cooper. High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nat. Commun.*, 9(1):2849, 2018.
- [627] J. Lehman, J. Clune, D. Misevic, C. Adami, J. Beaulieu, P.J. Bentley, S. Bernard, G. Belson, D.M. Bryson, N. Cheney, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *arXiv preprint arXiv:1803.03453*, 2018.
- [628] R. Gómez-Bombarelli. Reaction: The near future of artificial intelligence in materials discovery. *Chem*, 4(6):1189–1190, 2018.
- [629] G.Z. Yang, J. Bellingham, P.E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield, B.J. Nelson, B. Scassellati, M. Taddeo, R. Taylor, M. Veloso, Z.L. Wang, and R. Wood. The grand challenges of science robotics. *Sci. Robot.*, 3(14):eaar7650, 2018.
- [630] L. Zhang, R. Merrifield, A. Deguet, and G.Z. Yang. Powering the world’s robots-10 years of ROS. American Association for the Advancement of Science, 2017.
- [631] C.A. Nicolaou, I.A. Watson, H. Hu, and J. Wang. The proximal Lilly collection: Mapping, exploring and exploiting feasible chemical space. *J. Chem. Inf. Model.*, 56:1253, 2016.
- [632] H. Okamoto and K. Deuchi. Design of a robotic workstation for automated organic synthesis. *Lab. Robotics Automat.*, 12:2–11, 2000.
- [633] A. Adamo, R.L. Beingessner, M. Behnam, J. Chen, T.F. Jamison, K.F. Jensen, J.C.M. Monbaliu, A.S. Myerson, E.M. Revalor, D.R. Snead, T. Stelzer, N. Weeranoppanant, S.Y. Wong, and P. Zhang. On-demand continuous-flow production of pharmaceuticals in a compact, reconfigurable system. *Science*, 352:61, 2016.
- [634] J.L. Johnson, H.T. Wörden, and K. van Wijk. PLACE: An open-source Python package for laboratory automation, control, and experimentation. *J. Lab. Autom.*, 1:10–16, 2014.
- [635] M. Gronle, W. Lyda, M. Wilke, C. Kohler, and W. Osten. Itom: an open source metrology, automation, and data evaluation software. *Appl. Opt.*, 53:2974–2982, 2014.
- [636] M. Bates, A.J. Berliner, J. Lachoff, P.R. Jaschke, and E.S. Groban. Wet lab accelerator: A web-based application democratizing laboratory automation for synthetic biology. *ACS Synth. Biol.*, 6:167–171, 2017.

- [637] E. Whitehead, F. Rudolf, H.M. Kaltenbach, and J. Stelling. Automated planning enables complex protocols on liquid-handling robots. *ACS Synth. Biol.*, 7:922–932, 2018.
- [638] C. Houben and A.A. Lapkin. Automatic discovery and optimization of chemical processes. *Curr. Opin. Chem. Eng.*, 9:1, 2015.
- [639] R. Matsuda, M. Ishibashi, and Y. Takeda. Simplex optimization of reaction conditions with an automated system. *Chem. Pharm. Bull.*, 36:3512–3518, 1988.
- [640] J.M. Dixon and J.S. Lindsey. Performance of search algorithms in the examination of chemical reaction spaces with an automated chemistry workstation. *SLAS Technol.*, 9:364–374, 2004.
- [641] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, and A. Zeilinger. Automated search for new quantum experiments. *Phys. Rev. Lett.*, 116:090405, 2016.
- [642] T.P. Hughes. *American genesis: A century of invention and technological enthusiasm, 1870-1970*. University of Chicago Press, 2004.
- [643] L.A. Corkan and J.S. Lindsey. Experiment manager software for an automated chemistry workstation, including a scheduler for parallel experimentation. *Chemom. Intell. Lab. Syst.*, 17:47–74, 1992.
- [644] F. Häse. ChemOS: orchestrating self-driving laboratories. <https://github.com/aspuru-guzik-group/ChemOS>, 2018.
- [645] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A.Y. Ng. ROS: An open-source Robot Operating System. In *ICRA Workshop on Open Source Software*, volume 3, page 5, 2009.
- [646] A. Aspuru-Guzik and K. Persson. Materials Acceleration Platform: Accelerating advanced energy materials discovery by integrating high-throughput methods and artificial intelligence. *Mission Innovation Workshop*, 2018.
- [647] B. Cao, L.A. Adutwum, A.O. Oliynyk, E.J. Lubner, B.C. Olsen, A. Mar, and J.M. Buriak. How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. *ACS Nano*, 12(8):7434–7444, 2018.
- [648] X.D. Xiang, X. Sun, G. Briceno, Y. Lou, K.A. Wang, H. Chang, W.G. Wallace-Freedman, S.W. Chen, and P.G. Schultz. A combinatorial approach to materials discovery. *Science*, 268(5218):1738–1740, 1995.
- [649] H.S. Stein, D. Guevarra, P.F. Newhouse, E. Soedarmadji, and J.M. Gregoire. Machine learning of optical properties of materials—predicting spectra from images and images from spectra. *Chem. Sci.*, 10(1):47–55, 2019.
- [650] J. Li, Y. Lu, Y. Xu, C. Liu, Y. Tu, S. Ye, H. Liu, Y. Xie, H. Qian, and X. Zhu. AIR-Chem: Authentic intelligent robotics for chemistry. 122(46):9142–9148, 2018.

- [651] V. Gopalaswamy, R. Betti, J.P. Knauer, N. Luciani, D. Patel, K.M. Woo, A. Bose, I.V. Igumenshchev, E.M. Campbell, K.S. Anderson, et al. Tripled yield in direct-drive laser fusion through statistical modelling. *Nature*, 565(7741):581–586, 2019.
- [652] A. Milo, E.N. Bess, and M.S. Sigman. Interrogating selectivity in catalysis using molecular vibrations. *Nature*, 507(7491):210, 2014.
- [653] R.D. King, K.E. Whelan, F.M. Jones, P.G.K. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, and S.G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971):247, 2004.
- [654] L. Meng, Y. Zhang, X. Wan, C. Li, X. Zhang, Y. Wang, X. Ke, Z. Xiao, L. Ding, R. Xia, H.L. Yip, Y. Cao, and Y. Chen. Organic and solution-processed tandem solar cells with 17.3% efficiency. *Science*, 361(6407):1094–1098, 2018.
- [655] A.S.D. Sandanayaka, T. Matsushima, F. Bencheikh, S. Terakawa, W.J. Potscavage Jr, C. Qin, T. Fujihara, K. Goushi, J.C. Ribierre, and C. Adachi. Indication of current-injection lasing from an organic semiconductor. *Appl. Phys. Express*, 12(6):061010, 2019.
- [656] S. Jhulki and J.N. Moorthy. Small molecular hole-transporting materials (HTMs) in organic light-emitting diodes (OLEDs): structural diversity and classification. *J. Mater. Chem. C*, 6(31):8280–8325, 2018.
- [657] X. Yang, H. Wang, B. Cai, Z. Yu, and L. Sun. Progress in hole-transporting materials for perovskite solar cells. *J. Energy Chem.*, 27(3):650–672, 2018.
- [658] J. Burschka, F. Kessler, M.K. Nazeeruddin, and M. Grätzel. Co (III) complexes as p-dopants in solid-state dye-sensitized solar cells. *Chem. Mater.*, 25(15):2986–2990, 2013.
- [659] S. Wang, M. Sina, P. Parikh, T. Uekert, B. Shahbazian, A. Devaraj, and Y.S. Meng. Role of 4-tert-butylpyridine as a hole transport layer morphological controller in perovskite solar cells. *Nano Lett.*, 16(9):5594–5600, 2016.
- [660] T. Bu, L. Wu, X. Liu, X. Yang, P. Zhou, X. Yu, T. Qin, J. Shi, S. Wang, S. Li, Z. Ku, Y. Peng, F. Huang, Q. Meng, Y.B. Cheng, and J. Zhong. Synergic interface optimization with green solvent engineering in mixed perovskite solar cells. *Adv. Energy Mater.*, 7(20):1700576, 2017.
- [661] Y. Fang, X. Wang, Q. Wang, J. Huang, and T. Wu. Impact of annealing on spiro-OMeTAD and corresponding solid-state dye sensitized solar cells. *Phys. Status Solidi A*, 211(12):2809–2816, 2014.
- [662] P. Friederich, V. Meded, A. Poschlad, T. Neumann, V. Rodin, V. Stehr, F. Symalla, D. Danilov, G. Lüdemann, R.F. Fink, I. Kondov, F. von Wrochem, and W. Wenzel. Molecular origin of the charge carrier mobility in small molecule organic semiconductors. *Adv. Funct. Mater.*, 26(31):5757–5763, 2016.

- [663] R.E. Brandt, R.C. Kurchin, V. Steinmann, D. Kitchaev, C. Roat, S. Levenco, G. Ceder, T. Unold, and T. Buonassisi. Rapid photovoltaic device characterization through bayesian parameter estimation. *Joule*, 1(4):843–856, 2017.
- [664] T.H. Schloemer, J.A. Christians, J.M. Luther, and A. Sellinger. Doping strategies for small molecule organic hole-transport materials: impacts on perovskite solar cell performance and stability. *Chem. Sci.*, 10:1904–1935, 2019.
- [665] Y. Cui, H. Yao, J. Zhang, T. Zhang, Y. Wang, L. Hong, K. Xian, B. Xu, S. Zhang, J. Peng, Z. Wei, F. Gao, and J. Hou. Over 16% efficiency organic photovoltaic cells enabled by a chlorinated acceptor with increased open-circuit voltages. *Nat. Commun.*, 10(1):2515, 2019.
- [666] X. Liu, Y. Yan, Y. Yao, and Z. Liang. Ternary blend strategy for achieving high-efficiency organic solar cells with nonfullerene acceptors involved. *Adv. Funct. Mater.*, 28(29):1802004, 2018.
- [667] K. Li, Y. Wu, Y. Tang, M.A. Pan, W. Ma, H. Fu, C. Zhan, and J. Yao. Ternary blended fullerene-free polymer solar cells with 16.5% efficiency enabled with a higher-LUMO-level acceptor to improve film morphology. *Adv. Energy Mater.*, 9(33):1901728, 2019.
- [668] R.A. Potyrailo, B.J. Chisholm, W.G. Morris, J.N. Cawse, W.P. Flanagan, L. Hassib, C.A. Molaison, K. Ezbiansky, G. Medford, and H. Reitz. Development of combinatorial chemistry methods for coatings: High-throughput adhesion evaluation and scale-up of combinatorial leads. *J. Comb. Chem.*, 5(4):472–478, 2003.
- [669] D.G. Anderson, S. Levenberg, and R. Langer. Nanoliter-scale synthesis of arrayed biomaterials and application to human embryonic stem cells. *Nat. Biotechnol.*, 22(7):863, 2004.
- [670] A. Jaklenec, A.C. Anselmo, J. Hong, A.J. Vegas, M. Kozminsky, R. Langer, P.T. Hammond, and D.G. Anderson. High throughput layer-by-layer films for extracting film forming parameters and modulating film interactions with cells. *ACS Appl. Mater. Interfaces*, 8(3):2255–2261, 2016.
- [671] X. Rodríguez-Martínez, A. Sánchez-Díaz, G. Liu, M.A. Niño, J. Cabanillas-Gonzalez, and M. Campoy-Quiles. Combinatorial optimization of evaporated bilayer small molecule organic solar cells through orthogonal thickness gradients. *Org. Electron.*, 59:288–292, 2018.
- [672] A. Sánchez-Díaz, X. Rodríguez-Martínez, L. Córcoles-Guija, G. Mora-Martín, and M. Campoy-Quiles. High-throughput multiparametric screening of solution processed bulk heterojunction solar cells. *Adv. Energy Mater.*, 4(10):1700477, 2018.
- [673] X. Du, T. Heumueller, W. Gruber, A. Classen, T. Unruh, N. Li, and C.J. Brabec. Efficient polymer solar cells based on non-fullerene acceptors with potential device lifetime approaching 10 years. *Joule*, 3(1):215–226, 2019.

- [674] A. Distler, P. Kutka, T. Sauermann, H.J. Egelhaaf, D.M. Guldi, D. Di Nuzzo, S.C.J. Meskers, and R.A.J. Janssen. Effect of PCBM on the photodegradation kinetics of polymers for organic photovoltaics. *Chem. Mater.*, 24(22):4397–4405, 2012.
- [675] E.T. Hoke, I.T. Sachs-Quintana, M.T. Lloyd, I. Kauvar, W.R. Mateker, A.M. Nardes, C.H. Peters, N. Kopidakis, and M.D. McGehee. The role of electron affinity in determining whether fullerenes catalyze or inhibit photooxidation of polymers for solar cells. *Adv. Energy Mater.*, 2(11):1351–1357, 2012.
- [676] M. Salvador, N. Gasparini, J.D. Perea, S.H. Paleti, A. Distler, L.N. Inasaridze, P.A. Troshin, L. Lüer, H.J. Egelhaaf, and C.J. Brabec. Suppressing photooxidation of conjugated polymers and their blends with fullerenes through nickel chelates. *Energy Environ. Sci.*, 10(9):2005–2016, 2017.
- [677] C. Zhang, T. Heumueller, S. Leon, W. Gruber, K. Burlafinger, X. Tang, J.D. Perea, I. Wabra, A. Hirsch, T. Unruh, N. Li, and C.J. Brabec. A top-down strategy identifying molecular phase stabilizers to overcome microstructure instabilities in organic solar cells. *Energy Environ. Sci.*, 12(3):1078–1087, 2019.
- [678] C. Xie, X. Tang, M. Berlinghof, S. Langner, S. Chen, A. Späth, N. Li, R.H. Fink, T. Unruh, and C.J. Brabec. Robot-based high-throughput engineering of alcoholic polymer: Fullerene nanoparticle inks for an eco-friendly processing of organic solar cells. *ACS Appl. Mater. Interfaces*, 10(27):23225–23234, 2018.

Glossary

AI (artificial intelligence)

Artificial intelligence is a field of scientific research at the interface of applied mathematics and computer science. Research in artificial intelligence focuses on the development of artificial agents to solve complex tasks through statistical methods, optimization, or other techniques.

API (application program interface)

An application program interface defines a set of distinct computing methods and routines to interact with a software system such as operating systems, database systems, or software libraries.

ASF (achievement-scalarizing function)

Achievement scalarizing functions span a set of strategies to enable multi-objective optimization. An achievement scalarizing function constructs a scalar merit from a set of several optimization objectives following provided preference information, such that the optimal merit values correspond to Pareto optimal solutions to the multi-objective optimization task. In this spirit, achievement scalarizing functions constitute an *a priori* approach to multi-objective optimization.

BEC (Bose-Einstein condensate)

A Bose-Einstein condensate is a state of matter consisting of a gas of bosons at low densities and low temperatures. Under these environmental conditions, most of the bosons occupy their lowest quantum states, such that typically microscopic quantum phenomena including wave function interference can be observed on the macroscopic scale.

BKDE (Bayesian kernel density estimation)

Kernel density estimation describes a non-parametric approach to determine the probability distribution of a random variable. Bayesian kernel density estimation assumes a prior probability distribution which is refined based on collected samples using Bayes' theorem (see Sec. 2.2.2).

BNN (Bayesian neural network)

A Bayesian neural network is a machine learning model that shares the general architecture of a traditional neural network. Its fundamental building blocks are neurons that are organized in a directed graph and, given an input, generate an output based on a non-linear transformation and an intrinsic set of weights and biases. The weights and biases of neurons in a Bayesian neural network, however, are modeled as random variables following specific probability distributions. A Bayesian neural network is trained by updating the probability distributions via Bayesian inference (see Sec. 2.2.2) on the training data.

CASSCF (complete active space self-consistent field)

The complete active space self-consistent field method is a multiconfigurational approach for the generation of qualitatively correct reference states of molecules. A complete active space self-consistent field calculation varies the coefficients of both the determinants and the basis functions in the molecular orbitals to obtain the total electronic wavefunction at the lowest energy configuration. As such, the linear combination of configuration state functions includes all that arise from a particular number of electrons in a particular number of orbitals.

CMA-ES (covariance matrix adaptation evolution strategy)

The covariance-matrix-adaptation evolution-strategy constitutes a stochastic, derivative-free optimization strategy for the global optimization of black-box functions. It is based on informed updates of the covariance matrix of a multivariate normal distribution from which candidate solutions to the optimization problem are sampled.

CNN (convolutional neural network)

Convolutional neural networks are a special class of deep neural networks commonly applied in image recognition. They form regularized variants of fully connected multi-layer perceptrons which are particularly suited to identify hierarchical patterns in data in tensor representations due to a shared-weights architecture and translation-invariant characteristics.

DBMS (database management system)

A database management system is a software for the creation and management of databases. Database management systems serve as the interface between an end-user or application program and the database. To that end, database management systems implement the basic operations on a database, including methods to create, read, update and delete database entries.

DFT (density functional theory)

Density functional theory describes a computational approach to determine the quantum mechanical ground state of a many-electron system based on the spatially dependent electron density. Density functional theory has emerged as a popular approach in physics, chemistry, and materials science to calculate fundamental properties of molecules and solids from their electronic structures.

DoE (design of experiments)

Design of experiments is a broadly defined term that generally comprises all strategies to conceive comprehensive tests that probe a given hypothesis in an experimentation setting. More specifically, design of experiments probes the influence of controllable variables (*factors*) on measurable responses of interest. Common approaches for design of experiments suggest experimental variables which mutually differ in one factor, while keeping others constant.

EET (excitation energy transfer)

Electronic excitation energy transfer can occur between two molecules with overlapping emission and absorption bands. If one molecule, usually referred to as the excitation energy donor, *D*, has been electronically excited and the other, referred to as the excitation energy acceptor, *A*, is in its ground state, then the Coulomb interaction between the two molecules can induce a reaction where the donor molecule is de-excited, the excitation energy is transferred, and the acceptor molecule is excited. Excitation energy transfer thus constitutes a spontaneous photon-less energy transfer.

FF (force field)

Force fields in the context of molecular modeling describe the functional form and parameters to calculate the potential energy of a many-particle system. Common studies using force fields focus on the computation of the dynamical behavior at a biomolecular system at the atomic level *via* molecular mechanics or molecular dynamics simulations. Force fields thus collectively comprise a set of atomic energy functions and interatomic potentials.

FIFO (first-in, first-out)

The first-in, first-out principle for organizing a data buffer describes the working mechanism of a queue of tasks. Tasks that are scheduled earlier are also executed earlier, such that first-in, first-out conserves the chronological order in which the tasks have been scheduled.

FMO (Fenna-Matthews-Olson)

The Fenna-Matthews-Olson complex is a pigment-protein complex commonly found in green sulfur bacteria and the first pigment-protein complex whose atomistic structure was resolved by X-ray spectroscopy (see Sec. 2.1). The Fenna-Matthews-Olson complex is crucial to the photosynthetic processes in green sulfur bacteria, specifically for the excitation energy transfer from the light-harvesting chromosomes to the reaction center. It consists of three identical monomers, each of which contains eight bacteriochlorophylls.

GAFF (generalized Amber force field)

The general Amber force field is a special case of a force field which is designed for rational drug design. Parameters in the generalized Amber force field cover most of the organic chemical space with atomic charges computed based on HF/6-31* RESP.

GP (Gaussian process)

A Gaussian process is a stochastic process where a collection of random variables is indexed by a time or a space variable. Any finite collection of these random variables follows a multivariate normal distribution.

GUI (graphical user interface)

Graphical user interfaces provide the user of a computer software with graphical icons and visual indicators to facilitate the usage and control of that software.

HEOM (hierarchical equations of motion)

The hierarchical equations of motion formalism is a non-perturbative approach to studying the time evolution of open quantum systems. This approach is applicable even at low temperatures where quantum effects cannot be neglected. To that end, the hierarchical equations of motion correctly account for all system-bath interactions and non-Markovian noise correlations when computing the time evolution of the density matrix of the considered system.

HOIP (hybrid organic-inorganic perovskite)

Hybrid organic-inorganic perovskites describe a class of materials composed of inorganic cations and anions in combination with organic compounds arranged in a specific type of crystal structure. They have emerged as promising semiconductor materials for light-harvesting applications and light-emitting devices.

HOPS (hierarchy of pure states)

The hierarchy of pure states constitutes an approach to solving the dynamics of open quantum systems with non-Markovian structured environments. The hierarchy of pure states obtains the time evolution of the density operator of the system as an ensemble average.

HPC (high-performance computing)

High-performance computing describes the use of parallel processing techniques to solve complex computational problems on distributed computing architectures.

HPLC (high-performance liquid chromatography)

High-performance liquid chromatography is a technique in analytical chemistry to separate, identify, and quantify the components of an unknown mixture of chemicals. The components of the unknown mixture are separated by passing the sample mixture through a column filled with an adsorbent material, which modulates the flow rates of the individual compounds in the mixture.

HTE (high-throughput experimentation)

High-throughput experimentation constitutes an experimentation strategy which relies on the massive parallelization of common experimental tasks to simultaneously execute a large number of experiments at different conditions. To this end, high-throughput experimentation leverages automated experimentation platforms and has been pioneered especially for the discovery of novel drugs and functional materials.

HTM (hole-transport material)

Hole transport materials are critical components of perovskite solar cells with high hole mobilities to facilitate the spatial separation of positive and negative charges.

HT (high-throughput)

High-throughput processes include various procedures which are amenable to parallelization. Individual executions of a high-throughput process can be realized simulta-

neously, where the throughput measures the number of possible executions in a given time.

IR (infrared spectroscopy)

Infrared spectroscopy describes an analytic tool commonly used to study the interaction of electromagnetic radiation with matter, specifically to analyze the absorption, emission, and reflectance of radiation in the infrared spectral regions.

MAD (mean absolute deviation)

Given two sequences of length n , $\alpha = (\alpha_i)_{i=1}^n$ and $\beta = (\beta_i)_{i=1}^n$, in a k -dimensional space, $\forall i \in [1, n] : \alpha_i, \beta_i \in \mathbb{R}^k$, the mean absolute deviation between these sequences, $\text{MAD}(\alpha, \beta)$, measures the expected deviation of the absolute difference between two associated sequences elements, *i.e.*,

$$\text{MAD}(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^n |\alpha_i - \beta_i|.$$

MD (molecular dynamics)

Molecular dynamics describes an approach to computer simulations for molecular modeling which relies on inter-atomic interactions and Newtonian mechanics to compute the time-evolution of molecular many-particle systems.

MLP (multi-layer perceptron)

Multi-layer perceptrons are a class of fully connected feedforward neural networks that consist of at least three layers of nodes. Each node in the layers constitutes an independent perceptron with an intrinsic threshold activation.

ML (machine learning)

Machine learning is a subfield of artificial intelligence, aiming to identify patterns in presented data to extract general rules to solve a given task. Importantly, the rules are *learned* from the presented data and not explicitly provided by the programmer.

NLP (natural language processing)

Natural language processing is a subfield of computer science concerned with computer enabled speech recognition, natural language understanding, and natural language generation. Artificial intelligence methods are frequently used to approach problems in natural language processing.

OPV (organic photovoltaic)

Organic photovoltaic solar cells, also referred to as organic solar cells, aim to provide low-cost solutions to artificial light-harvesting based on Earth-abundant materials.

OSC (organic solar cell)

Organic solar cells constitute light-harvesting devices based on phase-separated mixtures of two or more organic materials in bulk-heterojunction architectures to absorb

sunlight and generating an electron-hole pair at the interface of two (or more) organic materials.

PCA (principal component analysis)

Principal component analysis constitutes an unsupervised machine learning strategy that projects a set of observations of possibly correlated variables onto a set of linearly uncorrelated variables *via* an orthogonal transformation. As such, principal component analysis identifies the linear combinations of input variables with the highest variance across the dataset.

PCE (power conversion efficiency)

The power conversion efficiency is a critical characteristic of a solar cell's ability to convert light into electrical energy. It is measured as the ratio of the electrical power output and the incident light power.

PEDOT:PSS (poly(3,4-ethylenedioxythiophene) polystyrene sulfonate)

Poly(3,4-ethylenedioxythiophene) polystyrene sulfonate is a polymer mixture of two ionomers. While PEDOT carries positive charges, polystyrene sulfonate is negatively charged. Together, the two constituents form a macromolecular salt which yields high efficiencies among organic thermoelectric materials and is frequently applied as a transparent, conductive polymer with high ductility.

PEDOT (poly(3,4-ethylenedioxythiophene))

Poly(3,4-ethylenedioxythiophene) is a conjugated polymer based on polythiophene, which carries positive charges. This polymer exhibits remarkable optical and electronic properties which established its wide-spread use as an organic electronics material.

PSC (perovskite solar cell)

Perovskite solar cells constitute third-generation light-harvesting devices based on inorganic lead halide matrices containing inorganic or organic cations. Perovskite solar cells convert sunlight into electrical energy by the direct absorption and conversion of incoming photons into free electrons and holes.

PSO (particle swarm optimization)

Particle swarm optimization describes a nature-inspired, metaheuristic optimization algorithm for the identification of the global optimum of a bounded parameter domain. Although different flavors of particle swarm optimization exist, the behavior of the canonical particle swarm optimizer is based on the flocking behavior of gregarious animals. Given a population of candidate solutions, each candidate is iteratively under the influence of its own position and velocity, its local best-known position, and the best-known position in the entire search space. Particle swarm optimizers gained prominence due to their applicability to unsupervised, complex multi-dimensional problems over the last two decades.

PSS (polystyrene sulfonate)

Polystyrene sulfonates are a group of organic compounds with several applications. By themselves, they can be used to treat high blood potassium, while in combination with PEDOT, polystyrene sulfonate forms a widely used organic electronics material.

RC (reaction center)

The reaction center is a pigment-protein complex which facilitates essential steps of the photosynthetic activity of autotroph organisms such as plants, algae and some bacteria. The reaction center separates incoming electronic excitations, generated either from direct photon absorption or transferred as excitation energy via light-harvesting pigment-protein complexes, into an electron-hole pairs.

RF (random forest)

Random forests constitute an ensemble learning method for classification and regression tasks in machine learning. A random forest is constructed from a collection of decision trees trained on a given dataset, such that predictions of a random forest are computed as the average prediction of its individual decision trees for regression tasks and the mode of the classes for classification tasks.

RMSD (root mean square deviation)

Given two sequences of length n , $\alpha = (\alpha_i)_{i=1}^n$ and $\beta = (\beta_i)_{i=1}^n$, in a k -dimensional space, $\forall i \in [1, n] : \alpha_i, \beta_i \in \mathbb{R}^k$, the root mean square deviation between these sequences, $\text{RMSD}(\alpha, \beta)$, measures the expected deviation of the difference of the squares between two associated sequences elements, i.e.

$$\text{RMSD}(\alpha, \beta) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\alpha_i - \beta_i)^2}.$$

ROS (Robot Operating System)

The Robot Operating System constitutes a framework and a set of tools that provide the functionality of an operating system on a heterogeneous computing cluster, and is mostly targeted to robotic applications and peripheral hardware.

SCP (secure copy protocol)

The secure copy protocol defines a network communication standard based on the secure shell protocol to transfer computer files between a local host and a remote host or between two remote hosts.

TD-DFT (time dependent density functional theory)

Time-dependent density functional theory is a quantum chemical approach to compute the properties and behavior of many-body systems in the presence of time-dependent environmental influences such as electromagnetic fields.

TON (turnover number)

The turnover number measures the maximum number of chemical conversions a catalyst can facilitate before becoming inactivated.

UV/Vis (ultraviolet-visible spectroscopy)

Ultraviolet-visible spectroscopy describes an analytic tool commonly used to study the interaction of electromagnetic radiation with matter, specifically to analyze the absorption, emission, and reflectance of radiation in the ultraviolet and full, adjacent visible spectral regions.

VSEPR (valence shell electron pair repulsion)

Valence shell electron pair repulsion theory is a quantum chemical model to predict the geometry of molecules based on the number of valence electrons surrounding a central atom. The premise of this model is that the electrostatic repulsion between electron pairs surrounding an atom tends to be minimized, thus giving rise to the geometry of the molecule.

XPS (X-ray photoelectron spectroscopy)

X-ray photoelectron spectroscopy is a surface-sensitive quantitative spectroscopic technique which is used to measure the elemental composition of a sample material and its surface, typically up to depths of a few nano meters. X-ray photoelectron spectroscopy relies on the photoelectric effect which describes the emission of electrons from a material upon photon absorption.