



Statistical and Generative Models for Inference and Reasoning in Astrophysics and Beyond

Citation

Mudur, Nayantara Ganapati. 2025. Statistical and Generative Models for Inference and Reasoning in Astrophysics and Beyond. Doctoral Dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://dash.harvard.edu/handle/1/42743963>

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 (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.

Please share how this access benefits you. [Submit a story](#)

HARVARD
Kenneth C. Griffin



GRADUATE SCHOOL
OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Department of Physics
have examined a dissertation entitled

Statistical and Generative Models for Inference and Reasoning in Astrophysics
and Beyond

presented by Nayantara Ganapati Mudur

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature Douglas Finkbeiner

Typed name: Prof. Douglas Finkbeiner

Signature Cora Dvorkin

Typed name: Prof. Cora Dvorkin

Signature Michael Brenner

Typed name: Prof. Michael Brenner

Date: July 21, 2025

Statistical and Generative Models for Inference and Reasoning in Astrophysics and Beyond

A DISSERTATION PRESENTED
BY
NAYANTARA GANAPATI MUDUR
TO
THE DEPARTMENT OF PHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
PHYSICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
JULY 2025

©2025 – NAYANTARA GANAPATI MUDUR
ALL RIGHTS RESERVED.

Statistical and Generative Models for Inference and Reasoning in Astrophysics and Beyond

ABSTRACT

DESPITE CONSIDERABLE LEAPS in our understanding of the Universe, murmurs of discordance in the measurements of cosmological parameters and unanswered questions about the nature of dark matter and dark energy remain. Addressing these gaps will likely require advances in three complementary thrusts: observation, theory, and analysis methods to bridge the two. This thesis draws from tools across the evolving machine learning landscape to address problems in each of these domains. On the observational front, we used simple interpretable filters to perform a more pristine separation of stars from extragalactic interlopers to synthesize maps of Galactic extinction that possess lower correlations with the large scale structure of the Universe. We then turn our attention to two interrelated tasks in the domain of cosmological parameter estimation. Cosmological simulations play a key role in generating predictions from theory; however, they are often expensive. We demonstrate the utility of diffusion generative models as emulators of dark matter density fields conditional on cosmological parameters, and show that the same model can solve the inverse problem of constraining the cosmological parameters of an input field. Next, we present two investigations that examine the potential of large language models (LLMs) in accelerating the scientific research pipeline, in particular, scientific code generation. We first evaluate the ability of LLMs to

Thesis advisor: Professor Douglas P. Finkbeiner

Nayantara Ganapati Mudur

solve general numerical analysis problems. This thesis concludes by returning to cosmology, specifically, the theoretical frontier, where we introduce a framework to enable an LLM-driven agent to autonomously implement and propose alternate cosmologies.

Contents

TITLE PAGE	i
COPYRIGHT	ii
ABSTRACT	iii
TABLE OF CONTENTS	v
DEDICATION AND ACKNOWLEDGEMENTS	ix
LIST OF FIGURES	xx
LIST OF TABLES	xxii
INTRODUCTION	i
1 STELLAR-REDDENING-BASED EXTINCTION MAPS FOR COSMOLOGICAL APPLICATIONS	6
1.1 Introduction	7
1.2 Datasets	10
1.3 Stellar Posterior Inference and Selections	14
1.4 Reconstruction Schemes	20
1.5 Analysis	21
1.6 Discussion	36
1.7 Conclusion	40
1.8 Code and Data Availability	41
1.9 Acknowledgements	41
2 DIFFUSION MODELS FOR ASTROPHYSICAL FIELDS	46
2.1 Introduction	47
2.2 DDPM Background	48
2.3 Generative Models for Dark Matter Density Fields	49
2.4 A Generative Model for Interstellar Medium Fields	53
2.5 Conclusions and Future Work	54
2.6 Broader Impact	55

2.7	Acknowledgements	56
3	DIFFUSION-HMC: PARAMETER INFERENCE WITH DIFFUSION-MODEL-DRIVEN HAMILTONIAN MONTE CARLO	58
3.1	Introduction	59
3.2	Datasets, Architecture and Training	62
3.3	Summary Statistics	66
3.4	Parameter Inference	71
3.5	Conclusion	80
3.6	Code	83
3.7	Acknowledgements	83
4	FEABENCH: EVALUATING LANGUAGE MODELS ON MULTIPHYSICS REASONING ABILITY	85
4.1	Introduction	86
4.2	Datasets and Tasks	89
4.3	Approach	92
4.4	Evaluation Metrics	95
4.5	Results and Discussion	98
4.6	Conclusion and future directions	104
4.7	Acknowledgements	105
5	AN LLM-DRIVEN FRAMEWORK FOR COSMOLOGICAL MODEL-BUILDING AND EX- PLORATION	107
5.1	Introduction	108
5.2	Environment Design	111
5.3	Stages	113
5.4	Evaluation	115
5.5	Results	116
5.6	Conclusions	124
5.7	Acknowledgements	126
6	CONCLUSION	127
	APPENDIX A STELLAR-REDDENING-BASED EXTINCTION MAPS FOR COSMOLOG- ICAL APPLICATIONS	132
A.1	Details of Stellar Selections	132
A.2	Deriving the angular cross correlation estimator	137
	APPENDIX B DIFFUSION MODELS FOR ASTROPHYSICAL FIELDS	139
B.1	Experiments and Compute Time	139

APPENDIX C	DIFFUSION-HMC	141
C.1	Parameter Inference	141
C.2	Summary Statistics	148
APPENDIX D	FEABENCH: EVALUATING LANGUAGE MODELS ON MULTIPHYSICS	
	REASONING ABILITY	150
D.1	Related Work	150
D.2	Dataset Curation	152
D.3	Evaluation Details	155
D.4	Querying the COMSOL Multiphysics® API from Python	163
D.5	Agent Details	166
D.6	Qualitative Analysis	172
D.7	Can LLMs solve these problems in Python?	173
APPENDIX E	AN LLM-DRIVEN FRAMEWORK FOR COSMOLOGICAL MODEL-BUILDING	
	AND EXPLORATION	175
E.1	Cosmological Models	175
E.2	Environment Details	182
E.3	Pipeline Details	184
E.4	Experiment Results	186
REFERENCES		221

TO MY PARENTS, AND ANNAMMA, FOR NURTURING MY CURIOSITY.

Acknowledgments

I am immensely grateful to my advisor, Douglas Finkbeiner, for allowing me the academic and intellectual freedom to pursue questions that piqued my interest, and for the invaluable role he has played in my development as a researcher. I am inspired by his intellectual versatility, mathematical intuition, mentorship, and insights across a range of domains. I am indeed privileged to have pursued my PhD under his guidance.

I am also deeply grateful to my committee members, Cora Dvorkin and Michael Brenner, for their invaluable role in shaping this thesis, their willingness to meet with me despite their many commitments, and their feedback on my progress.

The mentors and collaborators I have known and worked with played an important role in my journey. Joshua Speagle, Catherine Zucker, Ziyang (Zoe) Zhu, Tanveer Karim, Ioana Zelko, and Jun Yin helped me navigate my early years. The work and intellectual insights of my amazing collaborators in academia – Core Francisco Park, Carolina Cuesta-Lazaro, Michael Toomey, Maria Tikhanovskaya, Eun-Ah Kim, Gregory Green, among others – shaped many of the works in this thesis.

I consider myself deeply fortunate to have had the opportunity to intern at Google Research and am profoundly grateful to Michael Brenner, Peter Norgaard, Subhashini Venugopalan, Yasaman Bahri, Jackson Cui, Paul Raccuglia, Stephan Hoyer, Gowoon Cheon, and Brian Williams for their

mentorship, teaching me to write better code, and facilitating my adventures down the rabbit hole of large language models and science.

Furthermore, I thank

- Jacob Barandes and Lisa Cacciabauda in particular, Peggy Herlihy, the Harvard Physics department and other admin members for making my journey so much smoother. There is no other institution (and city) I would rather have spent these formative years of my life at.
- Gabriel Kreiman and Cengiz Pehlevan – for their guidance during my sojourn in computational neuroscience in my first year.
- my undergraduate alma mater for its rigor and Anurag Sharma and Sumeet Agarwal at IIT for their wisdom and advising.
- Shirley Ho, Chris Stubbs and Julia Kempe– for their advice.

I am grateful to Salman Habib at the Argonne National Laboratory, where I first began working in astrophysics and cosmology, for taking a chance on me. I believe my trajectory would have looked quite different without his mentorship and guidance. In addition to Dr Habib, I am also thankful to Patricia Larsen and Katrin Heitmann and other members of the CPAC group for their guidance.

Turning to the Finkbeiner group, I could not have asked for a more amazing group of (concurrent) academic siblings: Justina Yang, Andrew Saydjari, Ana Sofia Uzsoy: I will miss the group meeting banter, fun (and productive) memories during Cape Code and research and computing discussions from which I have learned so much.

I am especially appreciative of:

- my Cambridge-ish area friends – Misha, Core, Justina, Helena, Will, Neeraj, Rylan, Saran, Anqi, Zechen, Mikail and the other MIT folk, and so many others – for coloring these years in their own unique ways with hilarious memories, smatterings of commiseration and conversations about everything.
- Misha – for being the perfect housemate, Justina – for being a wonderful officemate and Susanne – for helping me on this journey.
- my undergraduate friends – Apurva, Cheena, Aditi, Avaneep, the Dhananjays, Poorva, Siddhesh, and Aman – for patiently entertaining (and hopefully, also being regaled by) my monologues and offering useful perspectives and occasionally a shoulder. Apurva and Cheena – as an only child, you two are probably a first order approximation to siblings.

I thank my family, especially my aunt and uncle, Sharmila and Dilip Kandlur – for so generously hosting me and offering a brief home away from home – and my grandparents – for always inquiring about my education. Annamma, in particular, thank you for playing such a crucial role in my upbringing, and I am sad you aren't here today. I miss our many conversations and banter and our shared love of books. And most important of all, I am grateful to my parents: Meenal and Ganapati Mudur. I consider myself so blessed to have had parents like mine: without their emotional support and sacrifices, the freedom they gave me to chart my own course, their deep care and involvement in me, and most notably, sparking my interest in physics, I would never have started down this path.

List of figures

1.1	Redshift distribution of extragalactic reference objects in the Northern Galactic Hemisphere. These objects were used to compute the angular cross-correlation with LSS.	14
1.2	Decision boundaries for the PanSTARRS-WISE $r-i$, $z-W_I$ cut (upper) and the SDSS $u-g$, $g-r$ cut (lower).	18
1.3	Redshift distribution of extragalactic objects in the spectroscopically matched subset in the Northern Galactic Hemisphere before and after secondary selections are applied. Galaxies were mostly removed by the preliminary cuts.	24
1.4	Excess extinction in maps at the positions of reference extragalactic objects in redshift bins spanning $z=0-4$ for two null examples where no correlation should exist: a map of SFD rotated by 180° (left column), and Gaussian noise (right column). The top row computes errors using the bootstrapped error bars, shown as grayscale regions around the measured correlation. The lower row computes errors using the rotation-based error bars, shown as grayscale regions centered on zero. The three envelopes correspond to 1, 3, and 5σ	29

1.5 Excess extinction in maps at the positions of reference extragalactic objects in redshift bins $z = 0-4$ for emission-based maps (first row: SFD and the Planck 2016 GNILC dust emission map, **last row: the Lenz-Hensley-Doré map**), the **Bayestar** stellar-reddening based maps (second row: Bayestar17 and Bayestar19), our maps with FWHMs of 6.1' and 15' (third row). In the fourth row, we plot the signals for the Gaia TGE map and for the FWHM=6.1' map with only distance cuts applied. Without cuts intended to eliminate extragalactic objects, the contamination from QSOs exceeds 5σ for $1 < z < 2.5$. The sigma contours are derived by rotating the map by 100 angles in the range $30^\circ - 300^\circ$ and computing the root mean square error of the signals of the ensemble of rotated maps. For all maps, the signal was evaluated over the region $b > 50^\circ$.

..... 31

1.6 Histogram of SFD extinction values (top row). Two-dimensional histograms of the bias (left column) and uncertainty-normalized bias (right column) for Bayestar17, and our maps at FWHM=6.1' and FWHM=15' in the top, middle and bottom rows respectively. The zero point and uncertainty corrections have been applied here. The histograms are normalized such that the pixel values along the y axis for each value on the x axis (corresponding to a single SFD extinction bin) sum to 1. The dotted lines denote the 25th, 75th (blue) and 50th (black) bias percentiles for each SFD extinction bin.

..... 35

1.7	Standard deviations of the offsets of the map relative to SFD for all testbeds. Each datapoint is the standard deviation over the difference between the input map's values and SFD's values in $N_{\text{SIDE}}=2048$ pixels contained within a single $N_{\text{SIDE}}=32$ pixel.	36
1.8	Comparison of SFD, Bayestar17, the Gaia TGE map and the reconstructions at $\text{FWHM}=6.1'$ and $\text{FWHM}=15'$ over different patches of the sky, with the colorscale set by the 1st and 99th percentile of the pixel values of the $\text{FWHM}=6.1'$ map for the area under focus.	39
2.1	Four log cold dark matter mass density fields from the training data (top left) and from the sampled model (top right) at 128×128 . Four samples of dust from the training data (bottom left) and from the trained model (bottom right).	50
2.2	Power spectra, normalized pixel histograms and Minkowski functionals for 100 draws from the real fields and the trained models at 64×64 (left) and 128×128 (right). The envelopes in the power spectra and the Minkowski functionals' panels represent the standard deviation of the value of the statistic in each bin. The height of the bars in the pixel histogram is the mean height of the histogram bin and the error bar is the standard error over all 100 samples.	51
2.3	Ground truth, noised input, denoised image and its residual (denoised - truth), smoothed baseline and its residual (smoothed - truth).	54

3.1	Generated fields at different cosmologies. Upper Row: Power spectrum of the unlogged fields for five different validation parameters. The lines depict the mean power spectrum and the envelope indicates the 16th and 84th percentiles of the distribution. True simulations are shown in black, generated in blue. Lower Row: Mean and standard deviation envelopes for the density histograms of the log fields.	65
3.2	Generated ‘1P’ fields. Left column: Generated fields corresponding to the extreme values of each parameter for a single seed, with the other value held fixed at the fiducial value (0.3 for Ω_m and 0.8 for σ_8). Middle column: Power spectra of the generated fields for the same seed, for different values of each parameter, holding the other fixed. Right column: Mean and standard deviation for the ratio of the power spectra at the modified parameter value to the power spectra for the field at the fiducial parameter value (black) for 15 slices from the CAMELS dataset (solid) and 15 seeds for the generated fields from the diffusion model (dashed). The effect of modulating a parameter on the generated fields’ power spectra is consistent with that of the true fields. .	67
3.3	Generated ‘CV’ fields. Left: Power spectra of the 405 true and generated fields, with the mean, and 16 – 84th percentiles. Right: Correlation matrix of the power spectra of the true and generated fields. The lower triangular matrix corresponds to the correlation matrix of the true fields while the upper triangular matrix corresponds to the correlation matrix of the generated fields.	69

3.4	Generated ‘CV’ fields. Left: Ratio of the standard deviations of the generated fields to that of the true fields. Right: Standard deviations of the power spectra in each k bin for the generated and true fields.	69
3.5	Investigating the contribution of the timesteps used in the VLB sum. Left: One sigma contours of each $-2\Delta\ln L_t$ ’s <i>individual</i> contribution for different timesteps. The contours are all centered near the true parameter (black star) and become wider as t increases. Right: Mean, and 1 sigma predictions for a single field as a function of the number of timesteps used in the VLB sum optimized by the HMC. Reducing the number of timesteps used to compute the VLB does not significantly affect the diffusion model predictions.	73
3.6	Top: Comparison of the parameter estimates for a single field with the power spectrum based estimator and with the diffusion model likelihood. The star demarcates the true parameter corresponding to the input fields. The shaded contours demarcate the 68.27%, 95.45% and 99.73% confidence intervals. Lower Panel: Predicted parameter and ground truth parameter for 10 different fields. The error bars correspond to the 15.9 – 84.1 percentiles for the marginal probability distributions for each parameter for each input field.	76

3.7	First row: An example field corresponding to $\Omega_m, \sigma_8 = 0.33, 0.96$ with its perturbed counterparts, along with the ratio of the perturbed linear field’s power spectrum to that of the original field $\frac{Pk_{Noisy}}{Pk_{Original}}$. Second, Third rows: Mean and 15.8 – 84.1 percentile predictions using the Diffusion-HMC estimates (green) and the mean and sigma predictions obtained from the parameter inference network in Ref. ²²⁸, denoted by InfNet. The mean absolute percentage bias (of the true prediction) for each parameter is also indicated. The Diffusion-HMC constraints are more robust to perturbations to the input image.	78
4.1	An overview of the agent and environment design.	92
4.2	The Evaluator’s feedback (<i>left</i>) is passed to the ToolLookupAgent, that calls <i>tools</i> and returns their concatenated output. Violet, on the <i>left</i> indicates that the Verifier Feedback is returned by an LLM, and on the <i>right</i>, denotes the arguments chosen by the ToolLookupAgent to call the tools with. The numerical annotations highlight the correspondence between the errors and arguments.	94
4.3	Block-wise executability across 300 samples of code and Gemini-1.5-Pro. The physics block has the lowest executability. Error bars denote standard deviations.	101

5.1	Flow diagram depicting the input state for each stage (blue) the output artifacts (green) and the task that the agent has to perform. Once the agent has completed its work, the outputs are automatically evaluated on the basis of both binary milestones that track initial progress as well as numerical metrics. The evaluation box (brown) depicts a template evaluation report for the agent's 'submission'.	111
5.2	Target Artifact-based evaluation: The evaluation for Stage 2 compares the evolution of the relevant dark energy component (Ω_{fld} for <i>Thawing Quintessence</i>) and the matter power spectrum at two redshifts against their corresponding ground-truth values generated by a target implementation of the model. While the agent is able to generate executable code when provided with the <i>paper (right column)</i> , the numerical errors are much higher than when it is given a condensed, yet detailed <i>description (left column)</i>	118
5.3	A cosmological model proposed by the LLM.	121
A.1	Excess extinction in maps at the positions of reference extragalactic objects in redshift bins spanning $z=0-4$ for existing emission-based maps (top panel: SFD and the Planck 2016 (GNILC) dust emission map), existing stellar-reddening based maps (middle panel: Bayestar17 and Bayestar19) and our reconstructions with FWHMs of 6.1' and 15' (last panel). The sigma contours are obtained by bootstrapping over the reference sample objects at each redshift.	136
C.1	The scale / standard deviation of the cumulative noise in Equation 3.1 added to the image over different timesteps.	141

C.2	HMC chains for 300 samples, for the same field, starting at two different initial parameters: [0.11, 0.62] (beige) and [0.3, 0.8] (teal). The chains are well mixed by the cutoff we designate as our burn-in (100 samples), denoted by the dashed line.	144
C.3	Robustness comparison without a prior applied in the HMC case.	146
C.4	Robustness comparison with the first 2 noising timesteps dropped <i>and</i> without a prior applied in the HMC case.	147
C.5	Mean and standard error on the mean for the reduced χ_r^2 statistic of the power spectra of the 50 linear (left) and log (center) fields for each parameter relative to the 15 true fields' power spectra for that parameter, across 10 different parameters for each of 8 checkpoints. The dashed line demarcates the mean reduced χ_r^2 of the true fields using the other 14 true fields as the reference distribution (leave-one-out). Right: Mean and standard error on the mean of the p values of the distribution of the means of the 50 generated log fields relative to the distribution of the means of the 15 true fields for the same parameter. The p values are above 0.05 for all of the 8 checkpoints.	149
D.1	Prompt used to ask the LLM to evaluate the problem	159
D.2	Block-wise executability across the 300 initial samples of code (purple) with PhyDoc In-Context and in the best solution (green) across all problems. Error bars denote standard deviations.	166

D.3	Executability and number of errors over solutions returned by the ControllerAgent.	
	The scatter at o denotes the spread from the samples in the initial population and the	
	black line denotes the mean value for the metric at that state across all problems. Each	
	colored line demarcates a different problem.	171
E.1	Description for the <i>Thawing Quintessence</i> model.	178
E.2	Prompt entasking the agent to propose a novel cosmological model.	181
E.3	Claude Code’s implementation of the <i>Thawing Quintessence</i> Model	187

List of tables

1.1	Distribution of objects from the spectroscopically matched sample in the Southern Galactic Hemisphere in different classes as stellar selection cuts are incrementally applied. The number reports the number of objects in each class. The figure in brackets reports the fraction of objects belonging to a certain category after a particular choice of cuts is applied $\frac{n_{Class}^{AfterCuts}}{n_{Total}^{AfterCuts}}$. Each row in the table also has all the cuts in the rows preceding it, applied. (The percentages in each row sum to 100%).	24
1.2	Ablation Tests on the spectroscopically matched sample in the Southern Galactic Hemisphere: All cuts in the Secondary Stellar Selection EXCEPT a given cut are applied.	25
4.1	Summary of Evaluation Metrics	96
4.2	Code Metrics: Comparison on ModelSpecs across LLMs.	99
4.3	Physics Metrics: Comparison on ModelSpecs across LLMs.	99
4.4	Comparison across models on FEABench Large.	99
4.5	Code Metrics: Comparison across tasks, prompts and agents.	102

4.6	Physics Metrics: Comparison across tasks, prompts and agents.	102
5.1	Results for Target Artifact-based Evaluation Experiments	116
5.2	Results for LLM-proposed model experiments	120
A.1	Distribution of objects from the spectroscopically matched sample in the Northern Galactic Hemisphere in different classes as stellar selection cuts are incrementally applied. The number reports the number of objects in each class. The figure in brackets reports the fraction of objects belonging to a certain category after a particular choice of cuts is applied $\frac{n_{Class}^{AfterCuts}}{n_{Total}^{AfterCuts}}$. Each row in the table also includes all the cuts in the rows preceding it. (The percentages in each row sum to 100%).	135
A.2	Ablation Tests on the spectroscopically matched sample in the Northern Galactic Hemisphere: All cuts in the Secondary Stellar Selection EXCEPT a given cut are applied.	136
D.1	Code Similarity across LLMs.	156
D.2	Code Similarity across experiments.	156
E.1	Results for Repeated <i>Thawing Quintessence</i> runs.	195

Introduction

In *Catching crumbs from the table*⁴³, the writer Ted Chiang conjures up a world in which super-intelligent beings have surpassed humans in terms of their ability to contribute to research, relegating human scientists to focus on interpreting the findings of “metahuman science”. Such a prospect has remained within the realm of fiction in the two and a half decades since the story was published. However, today, some might consider the possibility of such a scenario less implausible in the near future. Whether or not this speculation is premature, it is undeniable that the fields of statistics and machine learning, and their interplay with the physical sciences has evolved dramatically in recent years, extending their use to novel modalities. This thesis will reflect this transformation primarily in the context of astrophysics.

Astrophysics and cosmology endeavor to decipher the oldest and grandest chronicle in science: that of the Universe itself. The ability to differentiate between cosmological models that could have produced a Universe such as ours would have profound consequences on our understanding of its origin, its evolution and its eventual fate. The ‘golden years’ of cosmology, beginning in the late 90s, heralded phenomenal leaps in our understanding of the cosmos. Bolstered by its ability to predict several observations, including imprints in the clustering of galaxies known as Baryon Acoustic

Oscillations (BAO) and properties of the Cosmic Microwave Background (CMB), the Lambda Cold Dark Matter (Λ CDM) model emerged as the canonical cosmological model. In recent years, however, discordant measurements between observations have pointed to potential chinks in the Λ CDM armor. Foremost among these is the ‘Hubble tension’, a roughly 5σ discrepancy between the inferred value of H_0 , or the expansion rate of the universe, between ‘early Universe (CMB)’ and ‘late Universe’ probes such as supernovae and Cepheids. What remains unknown is whether these ‘tensions’ are the result of subtle observational or systematic mismodeling or whether our acceptance of Λ CDM will ultimately prove to be a transient phase in our understanding of the cosmos.

Tackling these questions and furthering our understanding of dark matter and dark energy requires research on multiple complementary fronts: sharper observations and larger sample sizes, better analytical tools and systematics calibration and novel theoretical developments. Sky surveys and missions like the Dark Energy Spectroscopic Instrument (DESI)⁵⁰, the Vera C. Rubin Observatory’s Legacy Survey of Space and Time (LSST)¹⁰² and the impending Nancy Grace Roman Space Telescope⁹ will capture a deeper and sharper view of the Universe while simultaneously pushing the boundaries of our data analysis and imaging needs. On the theory front, particle physicists and theoretical cosmologists have spent years trying to devise alternate beyond Λ CDM models. This is a delicate endeavor: any compelling model must alleviate outstanding cosmological tensions, reproduce the successes of Λ CDM, and adhere to Bayesian and physics principles.

Statistics and machine learning tools play a key role in these efforts. Spectroscopic and photometric data and observations at many scales permit a linear, Gaussian or quasi-Gaussian treatment of summary statistics and distributions. Chapter 1 of this thesis explores this regime on the obser-

vational front, and applies a subset of these approaches to performing a stricter selection of objects from stellar catalogs. Interstellar dust, or ‘extinction’ acts like a mist that shrouds and distorts our view of the Large Scale Structure (LSS) and signals from the early Universe. The filtered catalog of stars is used to synthesize maps of interstellar extinction with lower levels of cosmologically relevant systematics.

Interstellar extinction maps derived from stars are often limited in resolution or by noise since they are derived from noisy point tracers of an underlying field. This prompted me to work on a broader question: are there ways to distill our expectations, derived from either simulations or other complimentary observations on what a distribution of fields “should” resemble?

Chapter 2 serves as a prelude in this direction by applying diffusion generative models to modeling two classes of astrophysical fields. Probabilistic generative models, of which diffusion generative models are a subset, have become recent additions to the statisticians’ toolbox, because of their ability to capture complex non-Gaussian distributions.

In Chapter 3, we turn our attention to the task of connecting theory to observation. Cosmological simulations have long been a crucial tool in this process. However, the expense of precisely simulating hydrodynamical forces at high resolution is prohibitive. I use conditional diffusion generative models to emulate cosmological fields conditional on physical parameters, and show that they also excel at the inverse task of constraining physical parameters using simulations.

Where Chapters 2 and 3 involved learning the statistical distributions of fields or *images*, Chapters 4 and 5 dwell on the hitherto less accessible modality of *language*. The recent phenomenal successes of large language models (LLMs) have now rendered the task of mimicking specialized target

distributions of language more possible, opening up new avenues of applications to modeling text and code. In Chapter 4, we first embark on a brief detour away from astrophysics to the question of code generation in a more general scientific context, and evaluate the ability of large language models to solve engineering modeling and numerical analysis problems by generating code.

This thesis concludes with a preliminary exploration of a tantalizing question: could *new theoretical ideas* emerge from the ‘LLM laboratory’? In Chapter 5, I build a framework that introduces an LLM-driven agent into the cosmological model-building and exploration pipeline, and I examine the agent’s capabilities to implement cosmological models and propose new theories.

The following chapter and associated appendix was originally published as Mudur, N., Park, C.F. and Finkbeiner, D.P., 2023. Stellar-reddening-based extinction maps for cosmological applications. *The Astrophysical Journal*, 949(2), p.47. DOI: 10.3847/1538-4357/acc32c (CC BY 4.0). *The placement of some figures / tables and their referencing text was updated.*

*“If it wasn’t for the mist we could see your home across the
bay,” said Gatsby.*

F. Scott Fitzgerald, *The Great Gatsby*

1

Stellar-reddening-based Extinction Maps for Cosmological Applications

COSMOLOGICAL SURVEYS MUST CORRECT THEIR OBSERVATIONS for the reddening of extragalactic objects by Galactic dust. Existing dust maps, however, have been found to have spatial correlations with the large-scale structure of the Universe. Errors in extinction maps can propagate systematic biases into samples of dereddened extragalactic objects and into cosmological measurements such as correlation functions between foreground lenses and background objects and the primordial non-gaussianity parameter f_{NL} . Emission-based maps are contaminated by the cosmic

infrared background, while maps inferred from stellar-reddenings suffer from imperfect removal of quasars and galaxies from stellar catalogs. Thus, stellar-reddening based maps using catalogs without extragalactic objects offer a promising path to making dust maps with minimal correlations with large-scale structure. We present two high-latitude integrated extinction maps based on stellar reddenings, with a point spread function of full-width half-maximum $6.1''$ and $15''$. We employ a strict selection of catalog objects to filter out galaxies and quasars and measure the spatial correlation of our extinction maps with extragalactic structure. Our galactic extinction maps have reduced spatial correlation with large scale structure relative to most existing stellar-reddening based and emission-based extinction maps.

1.1 INTRODUCTION

Upcoming Stage IV cosmological surveys, such as the Dark Energy Spectroscopic Instrument (DESI)⁵⁰ and the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST)¹⁰² will image tens of millions of galaxies, supernovae and quasi-stellar objects, and aim to constrain cosmological parameters at the sub-percent level. To maximally leverage the statistical power of these datasets, it is imperative to characterize and minimize possible sources of systematic errors, such as Galactic dust extinction.

The interstellar medium of the Milky Way is speckled with dust. Interstellar dust is valuable as a probe of the interstellar medium and in studies of molecular clouds and structures in the Galaxy. However, it also obscures and modifies our view of both stars and extragalactic objects due to ab-

sorption and scattering of light by the intervening dust column, an effect referred to as extinction.

Thus, dust maps are needed to subtract the contribution of dust extinction from the observed magnitude of a background object prior to performing subsequent cosmological analyses.

Dust maps can be broadly categorized into two classes – emission-based maps and extinction-based maps. Emission-based maps use thermal emission at far infrared wavelengths to derive the dust column density. The Schlegel-Finkbeiner-Davis map²⁰⁰ hereafter SFD combined the high angular resolution of the Infrared Astronomy Satellite IRAS;²⁴ map at $100\mu\text{m}$ with the Diffuse Infrared Background Experiment DIRBE;⁹¹ maps’ superior calibration and wavelength coverage to derive a dust map with an angular resolution of full-width half maximum = $6.1'$. SFD, along with the Planck thermal dust maps⁸ at 353, 545 and 857 GHz and the Planck dust optical depth map, have since been among the most common choices of maps for applications in cosmology. However, Ref.²³⁹ found that over regions of the sky with an average extinction of $A_{r,SFD} < 0.1$ mag, the number counts of galaxies increased and their average post-extinction correction color was bluer with increasing extinction – contrary to the effect physically expected due to the obscuration of dust. They posited that such an effect could be explained by localized fluctuations due to far infrared emission from background galaxies in the SFD extinction map. Furthermore, Ref.⁴⁵ found that out of ten emission and extinction-based maps, all except the HI-based Ref.¹²⁵ map possess significant correlations with extragalactic structure. In emission-based maps, this arises from imperfect removal of the cosmic infrared background, the cumulative emission of dusty star-forming galaxies^{120,68}.

Another dust mapping approach involves using inferred stellar extinctions, conditional on an inferred or fixed reddening law. The line-of-sight reddening to a star can be derived by combining

theoretical knowledge of stellar types and intrinsic magnitudes with observed photometry from surveys. Since each star traces the extinction up to its distance, stellar reddening based maps allow dust extinction to be reconstructed in three dimensions. For these maps, contamination arises from imperfect removal of galaxies and quasi-stellar objects (QSOs) from stellar catalogs, leading to systematic overestimation and underestimation of extinction at their positions.

Correlations of galactic extinction maps with large-scale structure are concerning because when dust maps are used to deredden extragalactic objects, they imprint biases on the objects' dereddened magnitudes, as in Ref.²³⁹, which can further affect downstream analyses. Ref.⁴⁵ highlight secondary effects such as a bias on cosmological parameter constraints derived from luminosity distance measurements of dereddened Type Ia Supernovae, and on overdensities and two-point correlation functions estimated from magnitude-limited samples of extragalactic objects that can further propagate into correlations between foreground lenses and background objects. These lensing-induced correlation functions e.g.,^{138,75} are important probes of the galaxy bias, or the relation between foreground galaxy overdensities and matter overdensities⁶⁷. Other recent work^{20,115} has also shown that dust extinction can be a significant source of systematics in other cosmological measurements. Since stellar-extinction based maps should only be contaminated by the presence of stray extragalactic objects in stellar catalogs, in principle, maps built using stellar catalogs that have these stray galaxies and QSOs removed, should offer a promising path to generating maps without correlations with large-scale structure.

In this paper, we describe a set of two stellar reddening-based maps with a well-defined point-spread function. We take more care to eliminate galaxies and QSOs from our star sample, and an-

alyze the extent of correlation of the resulting maps with extragalactic structure. The document is organized as follows: Section 1.2 describes the datasets used in the inference and analysis process, Section 1.3 describes the stellar selections we use to eliminate galaxies and quasars. Section 1.4 describes the steps used in our reconstruction. Section 1.5 examines the performance of the stellar selections on a spectroscopically-matched test sample, measures the spatial correlation with large scale structure of the maps we make in addition to other dust maps in the literature and assesses the extent of noise in different parts of the sky. Section 1.6 discusses potential uses of the maps and future work.

1.2 DATASETS

In this section we describe three datasets used in this work: 1. the star sample, including data used to remove galaxies and QSOs; 2. the spectroscopically matched sample used to tune those cuts; and 3. the extragalactic objects used as tracers of large-scale structure in Section 1.5.

1.2.1 THE STAR SAMPLE

Stellar-extinction maps are built using inferred per-star posteriors on reddening and distance modulus for objects from stellar catalogs. For each star, we constrain the posterior using the Bayestar stellar inference framework⁸¹, inputting magnitudes from Pan-STARRS1 and 2MASS along with Gaia parallaxes.

The Panoramic Survey Telescope and Rapid Response System (Pan-STARRS1) is a ground-based sky survey using the eponymous 1.8-meter diameter telescope on the island of Maui, in Hawaii^{38,73}.

The first data release, the 3π Steradian Survey, imaged three quarters of the sky north of $\delta = -30^\circ$ in 5 broadband filters denoted by $g_{P1}, r_{P1}, i_{P1}, z_{P1}, y_{P1}$ – with mean wavelengths of 490nm, 624nm, 756nm, 869nm and 964nm – from May 2010 through March 2014. Sources are detected and processed by the Pan-STARRS Image Processing Pipeline PSPHOT to generate photometric and astrometric measurements^{131,133,235}. Photometric calibration was performed using the ‘ubercalibration’ approach^{156,72}, that solves for photometric parameters by minimizing the variance of the flux from repeated measurements of the same stars^{197,132}, resulting in a precision of around 10 mmag in all 5 bands.

The Two Micron All Sky Survey 2MASS;^{58,211} was a ground-based full-sky survey that scanned the sky in three bands in the near-infrared – J, H, and K_s, centered at $1.25\mu\text{m}$, $1.65\mu\text{m}$ and $2.16\mu\text{m}$ – from June 1997 to February 2001 and produced a catalog of 471 million sources. The survey used two 1.3m telescopes, one at Mt. Hopkins, Arizona, and the other at the Cerro Tololo Inter-American Observatory (CTIO) in Chile.

The *Gaia* mission^{74,173} is a space-based mission that measures the three dimensional spatial and kinematic distribution of stars in the Galaxy. The satellite consists of three instruments – the astrometric instrument, that collects images in the broad G band (330-1050 nm), the blue (BP, 330-680nm) and red (RP, 630-1050nm) prism photometers and a radial velocity spectrometer. The astrometric instrument on the satellite consists of two identical three-mirror anastigmat (TMA) telescopes with apertures of $1.45\text{m} \times 0.50\text{m}$ with a shared focal plane, and uses the principle of scanning space astrometry²²⁴. The mission has measured the parallaxes and proper motions of over a billion stars, or around 1% of all the stars in the Milky Way. The Gaia Early Data Release 3³⁵ hereafter Gaia

EDR₃ consists of data collected in the first 34 months of the mission and includes the proper motions and parallaxes of over 1.8 billion sources brighter than $G=21$.

We use additional photometric bands from SDSS and WISE for each candidate star to reject galaxies and QSOs. These fields are not used in deriving the posteriors on distance modulus and reddening, or in our preliminary selections, but are used in the secondary selections (Section 1.3.2) to minimize contamination from extragalactic objects.

The Wide-field Infrared Survey Explorer WISE,²³⁷ mapped the sky in four infrared bands, centered at 3.4, 4.6, 12, and 22 μm – referred to as the W₁, W₂, W₃, and W₄ bands. The survey used a 40 cm cryogenically-cooled telescope. The 4-Band Cryogenic survey lasted from January 7 2010 to August 6 2010, followed by a 3-Band Cryo phase and the NEOWISE Post-Cryo mission¹³⁴ phase that ended on February 1, 2011. The AllWISE Data Release⁶⁰, that included the AllWISE Source Catalog, combined data from all three phases. The AllWISE Source Catalog consisted of the astrometry and photometry of over 747 million objects, with typically higher sensitivity in the W₁ and W₂ bands than previous WISE releases.

The Sloan Digital Sky Survey SDSS;²⁷ is an imaging and spectroscopic survey that started in 1998, and uses the 2.5 m Sloan Foundation Telescope at the Apache Point Observatory, in New Mexico⁸⁶. SDSS measured photometry in the u , g , r , i and z bands as well as the spectra of millions of stars, galaxies and quasars. We query SDSS Data Release 14⁴ $ugriz$ photometry for objects in our sample. We use the Large Survey Database framework¹⁰⁷ to cross match between datasets.

1.2.2 THE SPECTROSCOPICALLY MATCHED SUBSET

To train the color cuts described in Section 1.3.2 and the Appendix and to examine the efficacy of our secondary selections on a subset of labeled objects, we take a sample of stars from the catalog obtained after the preliminary set of selections for $b > 50^\circ$ (Section 3.1) and cross match to the SDSS Data Release 17 ‘SpecObj’ catalog⁵, requiring a detection in the latter. This catalog consists of all objects whose spectra were obtained, and includes the redshift estimated from the spectra as well as their spectroscopic class. The spectra of objects usually allow us to reliably identify them as quasars, galaxies or stars, since each of these classes typically have characteristic features that distinguish them. Thus this subset offers a testbed with true labels that allows us to examine how well a given set of stellar selections removes galaxies and quasars.

1.2.3 EXTRAGALACTIC REFERENCE OBJECTS

Following Ref.⁴⁵, we use four spectroscopic reference samples as tracers of the LSS: the Main galaxy sample from the New York University Value-Added Galaxy Catalog^{28,7,156}, BOSS LOWZ and CMASS LRGs¹⁷⁸ and the Data Release 14 Quasar catalog (DR14Q) from the extended Baryon Oscillation Spectroscopic Survey (SDSS-IV eBOSS)¹⁶². The Main Galaxy sample is similar to the SDSS Main galaxy sample described by Ref.²¹⁷ but slightly more inclusive owing to a fainter magnitude limit (18 vs 17.77) and the absence of the exclusion of small bright objects, among other changes. This sample mostly spans $z \lesssim 0.3$. The LOWZ and CMASS samples from the Baryon Oscillation Spectroscopic Survey (BOSS) span $z \lesssim 0.4$ and $0.4 \lesssim z \lesssim 0.7$. The DR14Q cata-

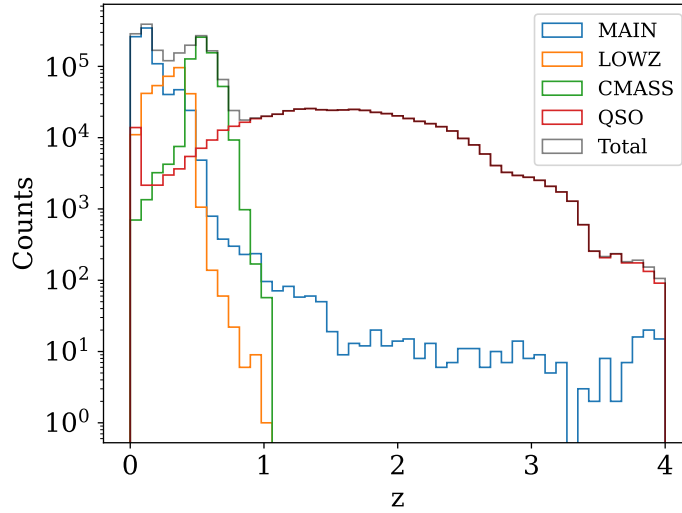


Figure 1.1: Redshift distribution of extragalactic reference objects in the Northern Galactic Hemisphere. These objects were used to compute the angular cross-correlation with LSS.

log compiled all spectroscopically-confirmed quasars from SDSS I through IV in the redshift range $0 \lesssim z \lesssim 5$. The distribution of this set of extragalactic reference objects is plotted in Figure 1.1, similar to Figure 5 in Ref. ⁴⁵.

1.3 STELLAR POSTERIOR INFERENCE AND SELECTIONS

1.3.1 PRELIMINARY SELECTIONS

Objects are crossmatched from the Pan-STARRS₁ catalog to the 2MASS catalog, and the Gaia EDR₃ catalog, requiring a detection in Gaia EDR₃. We additionally crossmatch to the AllWISE, and the SDSSDR₁₄ datasets. Note, all matches other than the one with Gaia are not exclusive and do not require a detection in the target survey. All survey crossmatches use a radius of 1 arcsecond, except the crossmatch to Gaia, where a radius of 0.5 arcseconds is used. The median proper motion

(Gaia EDR3 pm) of the sources is around 7 mas yr^{-1} , and is unlikely to be a problem for the cross-match criteria we use for most sources. A preliminary set of selections is made based on the quality of the data in PanSTARRS and 2MASS. The cuts on PS1 data require at least 2 good detections out of the 5 PS1 bands, on the basis of the `nmag_ok` parameter and that the difference between the reported PSF magnitudes and Aperture magnitudes be less than 0.1 in at least 2 bands. The procedure further requires ‘good detections’ in at least 4 out of 8 bands across both surveys and at least 2 good detections in PS1. The exact quality criteria that define ‘good detections’ for each survey are outlined in Appendix A.1.

The cuts most relevant to the question of separating extragalactic objects from true stars include the PSF-Aperture cut for PS1, and the `gal_contam` and `ext_key` based criteria for defining ‘good detections’ for 2MASS. The first of these selections is based on the observation that PSF photometry tends to underestimate the flux from extended objects like galaxies relative to the aperture photometry of extended objects e.g.,²¹⁷. The relevant 2MASS based criteria flag extended objects or those that lie within the $20 \text{ mag arcsecond}^{-2}$ elliptical isophote of an extended source. The specifics of these cuts are outlined in Appendix A.1.

This catalog of objects consists of magnitudes in eight bands (five for PS1 and three for 2MASS), as well as parallax information from Gaia EDR3. This information is used in deriving per star posteriors on distance modulus and reddening for each source.

Stellar Posterior Inference: We use the same stellar posterior inference framework as was used for the Bayestar19 dustmap, but with parallax measurements from Gaia EDR3. The stellar posterior inference framework only uses parallax information for ‘high fidelity’ Gaia detections. We

define high fidelity detections as those which have a Renormalized Unit Weight Error (RUWE) less than 1.4 and an ‘astrometric fidelity’ prediction greater than 0.5. The fidelity prediction for each source is obtained by querying the table generated by the astrometric fidelity classifier^{193,194}, a neural network that predicts the fidelity of a source conditioned on input features such as the RUWE, `astrometric_excess_noise`, parallax and proper motion measurements, among others.

The stellar posterior inference framework is described in Ref.⁸⁴, Ref.⁸⁵, and Section 2.1 and Appendix A in Ref.⁸¹, and summarized here. The observed magnitude of an object in any band $\hat{m}_b$ is modeled as the sum of the distance modulus, the intrinsic magnitude and the extinction in that band, with an additional noise term. The $\hat{\cdot}$ symbol is used to denote observable quantities, which are the magnitude in each band $\hat{m}_b$ and the parallax $\hat{\varpi}$.

$$\hat{m}_b \sim \mathcal{N}(m_{b,tb}, \hat{\sigma}_{m(b)}^2)$$

$$\text{where, } m_{b,tb} = M_b(\Theta) + \mu + A_b \quad (1.1)$$

$$\text{and, } A_b = R_b E$$

$$\implies \hat{m}_b - M_b(\Theta) \sim \mathcal{N}(\mu + R_b E, \hat{\sigma}_{m(b)}^2) \quad (1.2)$$

R_b is the component of the reddening vector $\vec{R}$ that converts the wavelength-independent reddening to the extinction in band b . The reddening vector used is the same as that in Ref.⁸¹, that uses the prescription of Ref.¹⁹⁸, with $R'_V = 3.3$ and $A_H/A_K = 1.55$ from Ref.¹⁰¹*. Since the mag-

*Ref.¹⁹⁸ define a parameter R'_V that is similar to the usual $R_V \equiv A_V/E_{B-V}$. A value of $R'_V = 3.3$ corresponds to $R_V = 3.1$, typical of dust in the diffuse ISM.

nitudes are linear in extinction and distance modulus, and the photometric errors approximately Gaussian, the maximum likelihood distribution of the distance modulus and extinction, given stellar type are Gaussian (Equation 1.2). The maximum likelihood distance moduli and extinctions are calculated for an ensemble of stellar types, and the likelihood of the data given the stellar type and the maximum likelihood values. This comprises a Gaussian Mixture Model where the components (i) correspond to each stellar type, with the means, weights and covariances given by:

$$\begin{aligned}
\text{Mean}_i &= [\mu_i, E_i] \\
w_i &= p(\hat{m}, \hat{\omega} | \Theta_i, \mu_i, E_i) p(\Theta_i, \mu_i, E_i) \\
C_i &= C = (A^T \Omega A)^{-1} \\
\text{where, } A &= [\vec{1}, \vec{R}]_{|\vec{m}| \times 2} \text{ and } \Omega = \mathcal{D}(\sigma_m^{-2})
\end{aligned} \tag{1.3}$$

The 16th, 50th, 84th percentiles of this distribution were computed and then converted to a median and standard deviation of the distance modulus and extinction.

Bayestar Stellar Models: The model described above requires the absolute magnitude in different bands as a function of stellar type $M_b(\Theta)$. The stellar type, Θ is parameterized by the absolute magnitude in the PS1 r band M_r , and the metallicity $[\text{Fe}/\text{H}]$. The prescription for deriving $M_b(\Theta)$ is described in detail in Ref. ^{85,81} and involves fitting a stellar locus in 7-D color space to a set of around 1 million stars in low reddening ($E(B - V)_{\text{SFD}} < 0.1$ mag) regions. This locus is then associated with an absolute magnitude and a metallicity using Ref. ¹⁰³ to yield an ensemble of stellar models $\{\vec{M}(M_r, [\text{Fe}/\text{H}])\}_i$.

In contrast to Bayestar19, we aim to make a 2-D map with cosmological applications in mind. A 2-D map contains far fewer degrees of freedom to constrain, so we can afford to be somewhat conservative in our selection of stars. We want a sample of high purity that lies behind most of the Milky Way dust.

1.3.2 SECONDARY SELECTIONS

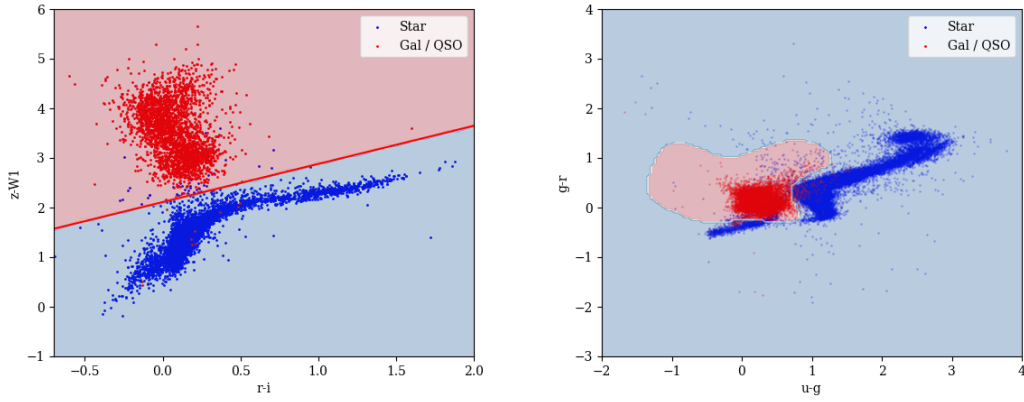


Figure 1.2: Decision boundaries for the PanSTARRS-WISE $r-i$, $z-W1$ cut (upper) and the SDSS $u-g$, $g-r$ cut (lower).

The combination of secondary stellar selections used in this paper is described below:

We retain in our sample only objects that pass the following cuts:

- $d_{50} \sin(b) > 400 \text{ pc}$, where d_{50} corresponds to the distance (in pc) derived from the median of the object's posterior on distance modulus.
- $\sigma_\mu < 1.5$ where $\sigma_\mu = (\mu_{84} - \mu_{16})/2$ and μ_P is the P^{th} percentile of the object's posterior on distance modulus.

- the object passes a ‘linear’ PanSTARRS-WISE color cut.
- the object passes a color cut using SDSS ugr .
- the object has a valid parallax measurement in Gaia EDR₃.
- the reduced chi-square $\chi_r^2 < 3$.

The first two cuts on distance modulus are based on our modelling assumptions. The third and fourth cuts, plotted in Figure 1.2, are color cuts whose decision boundaries are trained on a subset of objects with spectroscopic labels from SDSS DR17 ‘SpecObj’. The third cut is a color cut using WISE and PanSTARRS riz that retains objects that are identified as ‘stars’ by a support vector machine trained on their $m_r - m_i$ and $m_z - m_{W1}$ values. This cut effectively classifies all objects with $[m_z - m_{W1}] < b + A \times [m_r - m_i]$ as stars, where $b = 2.11, A = 0.77$. Unlike most stars, which typically have their radiation peak specifically in the optical, QSOs and galaxies have extended emission spectra, and relatively higher flux in the infrared. A similar boundary was also used by Ref.⁸¹ to assess the extent of quasar contamination, but not as a cut on objects. The fourth cut retains objects that are identified as ‘stars’ by a support vector machine that uses the objects’ $m_u - m_g$ and $m_g - m_r$ colors from the SDSS DR14 survey. This is motivated by the ultraviolet selection method: QSOs consist of Active Galactic Nuclei which often exhibit excess flux in the ultraviolet due to the blue bump. The training and implementation of these cuts is described in Appendix A2. The fifth cut, retains objects that possess a parallax detection in Gaia EDR₃ that isn’t a NaN. The sixth cut acts on the reduced chi-squared statistic of the object’s posteriors and excludes sources that are a poor fit to the Bayestar stellar models. We sequentially examine how effectively each cut removes extragalactic

objects in Section 1.5.1. While additional cuts on proper motion could theoretically increase the purity of objects included in the spectroscopically matched sample, we found that their addition did not seem to further reduce the angular cross-correlation of the resulting map, indicating that that information was possibly redundant with the other cuts described here.

1.4 RECONSTRUCTION SCHEMES

We choose to construct our maps at HEALPix⁷⁹ NSIDE=2048, and with a point spread function FWHM 6.1' and 15'. We derive a mean ($\text{Mean}(E[p])$) and a variance ($\text{Var}(E[p])$) extinction estimate for all pixels p in the map in the following fashion. For all sources that pass the selections in Section 1.3, sources within $5 l_{\text{PSF}}$ (where $l_{\text{PSF}} = \text{FWHM}/2.355$) of a given pixel p are filtered, and weighted by the expression below, where s denotes a source:

$$w_{\text{PSF}}[p, s] = \exp \left[\frac{-(\vec{r}_p - \vec{r}_s)^2}{2l_{\text{PSF}}^2} \right] \quad (1.4)$$

$$w_{\text{inv}}[p, s] = \frac{w_{\text{PSF}}}{\sigma_s^2 + \sigma_{\text{ref}}^2} \quad (1.5)$$

$$w_{\text{eff}}[p, s] = \frac{w_{\text{inv}}}{\sum_s w_{\text{inv}}} \quad (1.6)$$

$$\text{Mean}(E[p]) = \sum_s w_{\text{eff}}[p, s] * E_{50}[s] \quad (1.7)$$

$$\text{Var}(E[p]) = \sum_s w_{\text{eff}}^2[p, s] * \sigma_E^2[s] \quad (1.8)$$

$$\text{where, } \frac{1}{\sigma_{\text{ref}}^2 + \sigma_{1\%}^2} = \frac{10}{\sigma_{\text{ref}}^2 + \sigma_{99\%}^2} \quad (1.9)$$

σ_{ref} is set such that the inverse variance weighting factor assigned to a star with a posterior extinction uncertainty corresponding to the 1st percentile of stars which passed the cuts in Section 1.3 in the parent $N_{\text{SIDE}}=32$ pixel is 10 times larger than an equivalent star at the same distance but with a posterior extinction uncertainty corresponding to the 99th percentile. Since the maps’ reddenings are calibrated to SFD, we multiply the maps by 0.856, to convert to $E(B - V)$. As a postprocessing step, a constant is added to each map such that the resulting maps’ means match the mean of the extinction values predicted by SFD (also multiplied by 0.86) over the patch of the sky at $b > 60^\circ$. This postprocessing step does not affect the angular cross-correlation plots since the mean of the maps is subtracted as a preprocessing step before computing the estimate. The 0.86 factor multiplying SFD was derived in Ref. ¹⁹⁹ by recalibrating SFD with respect to the blue tip of the stellar locus using SDSS photometry. All comparisons with SFD in this work use this factor.

1.5 ANALYSIS

We use two figures of merit to assess the extent of extragalactic contamination in our maps.

The first test of how well our selection eliminates galaxies and quasars from the catalog sample is to examine a subset of objects which have spectroscopic labels, and assess the extent of contamination from extragalactic objects before and after applying the selection. The subset of objects with spectra is only a small fraction ($\sim 1\%$) of the full set of objects whose reddening posteriors are used in the reconstruction. This testbed allows us to directly examine the effect of adding or removing different cuts. The spectroscopically matched sample is not intended to be representative of the dis-

tribution of objects in our the full catalog. The spectroscopically matched sample is likely to contain more QSOs, since the distribution is reflective of the objects that were targeted to have their SDSS spectra measured.

The second, complementary test examines the spatial correlation of a dust map with LSS as traced by extragalactic objects with redshifts from the Sloan Digital Sky Survey (SDSS). For this we use the clustering-redshift based method⁴⁵, originally introduced in Ref.¹³⁹. This technique has been applied to infer the correlation with redshift of an unknown discrete sample or a continuous field with reference objects with known redshifts. This can be used to infer the redshift distribution of a sample of objects, or, as in Ref.⁴⁵, evaluate whether extinction patterns in maps correlate with the positions of objects in specific redshift bins. Our implementation of the measurement differs from their approach in certain respects, specifically in terms of how we measure the error on the cross-correlation signal, the smoothing scale used to preprocess the maps, the level of pixelization at which input maps are queried and the absence of an inverse variance weighting scheme.

We also examine the extent of deviations from SFD and the level of noise as a function of latitude.

1.5.1 PERFORMANCE ON A SPECTROSCOPICALLY MATCHED SAMPLE

To examine the efficacy of our secondary selections, we cross match stars from the catalog obtained after the preliminary set of selections (Section 1.3.1) for $b > 40^\circ$ to the SDSS Data Release 17 ‘SpecObj’ catalog, with a crossmatch radius of 1”. To ensure the spectroscopic labels are reliable we further impose the following quality cuts to select objects with ‘reliable’ spectra, and require

$r_{chi2} < 2$, $chi68p < 2$, and $sn_median_all > 10$. This yields a set of 210713 objects, of which 175597 (83%) are stars, 34817 (17%) are QSOs, and 299 are galaxies (0.14%). The low number of galaxies in the sample without any secondary selections applied can be attributed to the aperture minus PSF magnitude cut applied in the preliminary selections. The color cuts involving SVM models were trained on a subset of the spectroscopically matched sample from the Northern Galactic Hemisphere. After applying all cuts, 144750 stars, 42 QSOs and 49 galaxies remained in the sample. Figure 1.3 plots the normalized redshift distribution of objects in the spectroscopically matched sample before and after the cuts are applied. While high $z > 3$ objects account for a small fraction of extragalactic objects before making cuts, they account for a *relatively* larger fraction of objects that remain in the sample after applying the stellar selections. Tables 1.1 and 1.2 examine the behavior of the cuts on an independent ‘test’ set of spectroscopically matched objects with reliable spectra in the Southern Galactic Hemisphere, with $b < -40^\circ$. None of the objects in this set was used to train the color cuts. The equivalent tables for the Northern Galactic Hemisphere are in Appendix A.1.2. Before applying any secondary selections, this set contained 55474 (85%) stars, 9816 (15%) QSOs, and 87 (0.13%) galaxies. After applying the stellar selections, 47916 stars (99.91%), 30 galaxies (0.06%) and 14 QSOs (0.03%) are left. We examine the incremental effect of each cut one at a time on the sample, and perform ablation tests where the effect of excluding each cut are tested. The cuts on distance modulus have negligible effect on the purity of the sample, while the SDSS ugr and PanSTARRS-W1 cuts are particularly valuable in terms of removing QSOs. The Gaia EDR3 parallax detection cut had the weakest effect overall but was included because it appeared to reduce contamination from galaxies better than some of the other cuts as well as reduce noise in the result-

ing map.

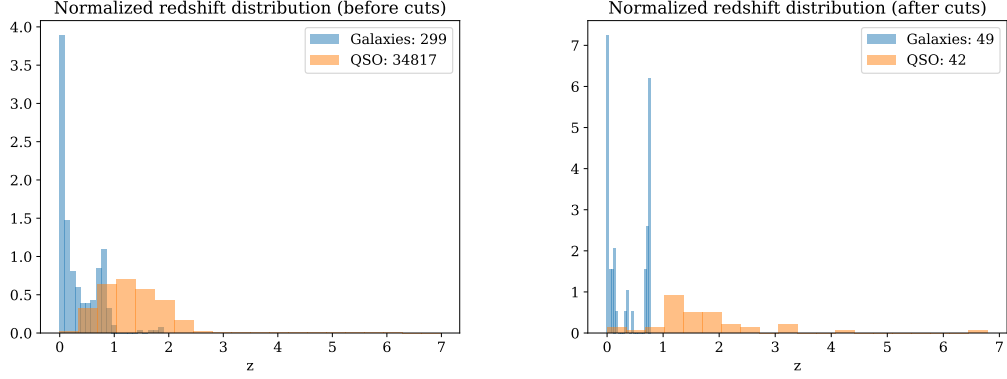


Figure 1.3: Redshift distribution of extragalactic objects in the spectroscopically matched subset in the Northern Galactic Hemisphere before and after secondary selections are applied. Galaxies were mostly removed by the preliminary cuts.

Table 1.1: Distribution of objects from the spectroscopically matched sample in the Southern Galactic Hemisphere in different classes as stellar selection cuts are incrementally applied. The number reports the number of objects in each class. The figure in brackets reports the fraction of objects belonging to a certain category after a particular choice of cuts is applied $\frac{n_{Class}^{AfterCuts}}{n_{Total}^{AfterCuts}}$. Each row in the table also has all the cuts in the rows preceding it, applied. (The percentages in each row sum to 100%).

Cuts	Stars: No./ Proportion (%)	QSOs: No./ Proportion (%)	Galaxies: No./ Proportion (%)
No Cuts Applied	55474 (84.85)	9816 (15.01)	87 (0.13)
+ $d_{50} \sin(b) > 400\text{pc}$, $\sigma_z < 1.5$	52383 (84.58)	9463 (15.28)	85 (0.14)
+WISE nondetection/ $(r - i, z - W1)$ cut	51835 (98.99)	470 (0.90)	57 (0.11)
+ EDR3 parallax detection	51641 (99.01)	470 (0.90)	47 (0.09)
+SDSS $(u - g, g - r)$ cut	51438 (99.84)	44 (0.08)	41 (0.08)
+ $\chi_r^2 < 3$	47916 (99.91)	14 (0.03)	30 (0.06)

1.5.2 MEASURING CORRELATIONS WITH LARGE SCALE STRUCTURE

We examine the angular cross correlation of extragalactic reference objects in different redshift bins with three existing emission-based maps, three existing extinction-based maps, and the new maps we construct, using the clustering-based redshift estimation technique.

Table 1.2: Ablation Tests on the spectroscopically matched sample in the Southern Galactic Hemisphere: All cuts in the Secondary Stellar Selection EXCEPT a given cut are applied.

Excluded Cut	Stars: Number	QSOs: Number	Galaxies: Number
No cuts excluded	47916	14	30
-WISE nondetection/ $(r - i, z - W1)$ cut	48179	36	31
-EDR ₃ parallax detection	48086	14	34
-SDSS $(u - g, g - r)$ cut	48047	69	32
$\chi_r^2 < 3$	51438	44	41

We briefly summarize dust mapping efforts below, and the maps we compare against. The Burstein-Heiles map³⁶ was one of the earliest attempts at dust mapping, and fit extinction as a function of galaxy counts and the column density of neutral hydrogen (HI). The Schlegel-Finkbeiner-Davis (SFD) map²⁰⁰, generated a map of dust emission at $100\mu\text{m}$ using measurements from IRAS and DIRBE⁹¹ with an angular resolution of $\text{FWHM} = 6.1'$, after subtracting zodiacal light emission from the DIRBE maps and removing striping artifacts and point sources from IRAS. This emission map was converted to optical depth at $100\mu\text{m}$ using a color temperature derived from the ratio of $100\mu\text{m}$ to $240\mu\text{m}$ emission in the DIRBE maps. The dust column density map was converted to optical reddening map by calibrating against the reddening of a sample of elliptical galaxies. The maps used by SFD have an estimate of the mean cosmic IR background subtracted, but not its anisotropy, leaving a residual correlation with large-scale structure. The extinction values from SFD are multiplied by the 0.86 factor derived by Ref.¹⁹⁹.

The Planck 2016 products (Ref.⁸) include dust emission maps at 353, 545 and 857 GHz. The inputs to the map-generation procedure were the Planck temperature full-mission sky maps from the Planck 2015 data release (PR2) (Ref.⁵¹, Ref.⁶) in nine frequencies ranging from 30 to 857 GHz, as

well as a combined temperature map at $100\ \mu\text{m}$ based on a combination of the IRIS map (Ref. ¹⁴²) and SFD. The map has a spatially varying effective beam size depending on the signal to noise ratio over the sky and has a FWHM of $5'$ over 65% of the sky but up to $15'-21'$ at the highest galactic latitudes. A dust extinction map is obtained by multiplying the dust optical density map at $353\ \text{GHz}$ by a factor derived from the correlation between the reddening of quasars and dust optical depth (Ref. ³).

The Lenz-Hensley-Doré¹²⁵ interstellar reddening map was derived from Galactic HI emission at low column densities ($N_{\text{HI}} < 4 \times 10^{20}\text{cm}^{-2}$) and line-of-sight velocities of less than 90kms^{-1} and has a resolution of $16.1'$. We replaced missing patches in the map with the mean extinction value at $b > 50^\circ$ before computing the angular cross-correlation. The missing pixels account for 2.5% of all pixels at NSIDE=2048 at $b > 50^\circ$.

In recent years, several stellar-reddening based maps have been produced, differing in terms of their statistical reconstruction methods, volume covered, extents of discretization and choices of stellar posteriors eg: ^{85,184,83,81,121,124}. The Bayestar17 map⁸³ used photometry from Pan-STARRS 1³⁸ and 2MASS²¹¹ to generate a stellar-reddening map in three dimensions for $\delta > -30^\circ$. This work used the reddening vector derived by Ref. ¹⁹⁸ from the measurements of the dust extinction curves of thousands of stars with spectra from the APOGEE spectroscopic survey¹³⁵ and ten photometric bands in the optical and infrared. The Bayestar19⁸¹ map broadly adopted the same approach to deriving a three dimensional dust distribution, but differed in some key aspects, including the addition of distance constraints in the form of parallaxes from the Gaia Data Release 2³⁴, a Gaussian Process based prior to regularize the spatial dust distribution and the discretization of the incremen-

tal extinction to multiples of 0.01 mag. The reddenings from the Bayestar maps were multiplied by a factor of 0.856 to convert to units of E_{B-V} . The Gaia Total Galactic Extinction map⁶³, described in Section 1.7, is another two-dimensional stellar reddening based map and used extinctions derived from Gaia DR3 data. We use the DUSTMAPS package to query these maps⁸⁰, and compute the angular cross-correlation of extragalactic objects in different redshift bins with excess reddening in each of these maps.

IMPLEMENTATION

Our implementation of the angular cross-correlation estimator is described below, and in Appendix A.2:

1. Each input map $E(\varphi)$ is queried at HEALPix N_{SIDE}=2048.
2. The input map is smoothed at $\sigma_{GAL} = 30'$ to identify correlations arising from specifically extragalactic signals. The dust density is, to first order, largely dependent on galactic latitude and to second order on filaments and structures in galactic dust. Thus, subtracting the input map smoothed at a scale σ_{GAL} would largely eliminate these dependencies while preserving local fluctuations arising specifically from correlations with extragalactic structure. The mean of this difference is further subtracted to yield $\delta E(\varphi)$ (Equation A.1).
3. A mask is defined that consists of the relevant area over which the cross-correlation signal is calculated. For the purpose of our analysis, our mask consists of the intersection of $b > 50^\circ$ and the area where there are extragalactic tracers with which we can compute the cross

correlation.

4. We then evaluate the cross-correlation of the processed map with the fractional overdensity of reference objects at a given redshift, in each angular bin upto a certain angular scale set by $\theta_{max} = 9.73'$ for each pixel (see Equation A.4). This gives us the angular cross-correlation signal in 34 redshift bins from $z = 0.029 - 4.048$ for the input map.
5. To estimate the error on this signal, arising from the intrinsic noise of the map, we rotate the map about the north Galactic pole to 100 angles in the range $30^\circ - 300^\circ$, and measure the root mean square deviation of the signal for the ensemble of rotated maps. In subsequent paragraphs, an x ‘sigma’ contour refers to x times the RMSE computed in this fashion. The intuition underlying this step is that there is no physical reason a point on the sky would possess small-scale correlations with another random point on the sky – such as that obtained by rotating the frame by an angle larger than some small angle. Thus, in theory the signal for the rotated maps should be 0 at all redshifts. Any deviations from 0 are thus a measure of the contribution of random correlations between fluctuations in the map and the distribution of the reference objects to the estimate of the extragalactic correlation signal. Bootstrapping over the reference objects in the sample for each redshift bin underestimates the error bars relative to this approach to estimating the error on the correlation signal, as demonstrated in Section 1.5.2.

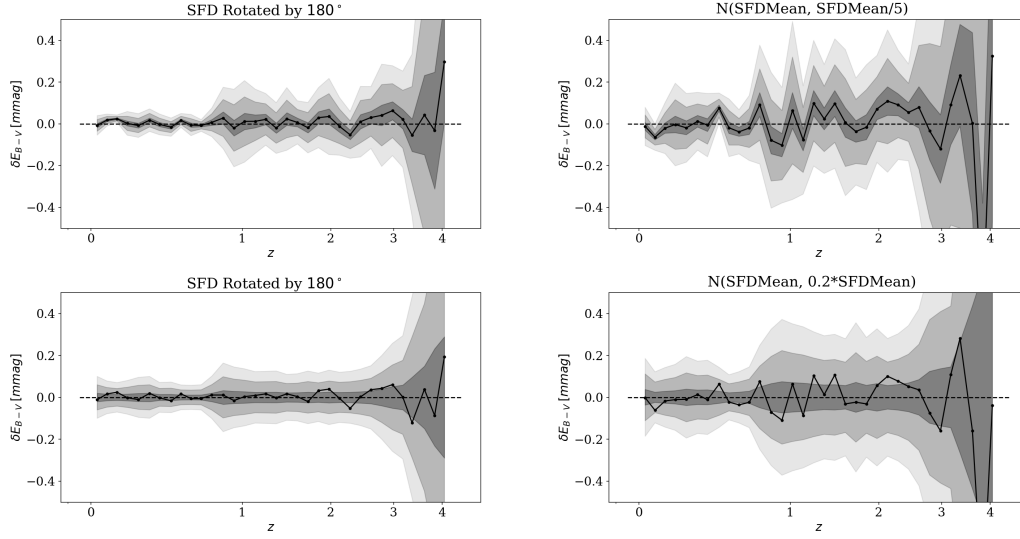


Figure 1.4: Excess extinction in maps at the positions of reference extragalactic objects in redshift bins spanning $z=0-4$ for two null examples where no correlation should exist: a map of SFD rotated by 180° (left column), and Gaussian noise (right column). The top row computes errors using the bootstrapped error bars, shown as grayscale regions around the measured correlation. The lower row computes errors using the rotation-based error bars, shown as grayscale regions centered on zero. The three envelopes correspond to 1, 3, and 5σ .

ERROR BARS

Figure 1.4 depicts the angular cross-correlation signal for two null examples – maps which we know should not possess any physical correlations with extragalactic structure. The null examples with the bootstrapped error bars, in the top panel seem to have more points that are around 5 sigmas off, such as the signal in the third redshift bin for SFD rotated by 180° . To derive the error bars for SFD rotated by 180° , we rotate the map to 100 angles between $21.8^\circ - 310.9^\circ$ and exclude rotations between $170^\circ - 190^\circ$, since the rotating the input by 180° returns the original map i.e. a map that does possess correlations. The level of signal measured using the rotation-based error bars does seem to lie largely within the 1 sigma contours. The differences in the choice of centering the error bars

arises from the fact that the rotation-based errors account for deviations of the signal from 0 while the bootstrapped errors, account for deviations of the signal from its measured value, since we are bootstrapping over possible reference samples. For the majority of our analysis, we use the rotation-based error bars. We additionally plot angular cross-correlations for all the dust maps we consider using the bootstrapping-based errors in the Appendix.

COMPARISONS

We use this estimator, and the rotation-based error bars, to assess LSS contamination of three emission based maps – SFD, GNILC and the Lenz-Hensley-Doré (LHD) map – three stellar reddening based maps – Bayestar17, Bayestar19 and the Gaia TGE map – and the maps we construct in Figure 1.5. A positive correlation *typically* indicates contamination from galaxies. In emission-based maps, this arises from imperfect removal of the cosmic infrared background, which leads to overestimated extinctions. Since it is foreseeable that the positions of galaxies in the reference samples are likely to correlate with sources of the cosmic infrared background (dusty star-forming galaxies), this leads to a higher cross-correlation signal at lower redshifts ($z < 1$). We find that SFD’s signal exceeds the 5 sigma contour at redshifts less than $z \sim 1$. The GNILC map has a lower LSS correlation, with the signal within the 5 sigma level, while LHD has the lowest correlation out of the three emission-based maps. As in Ref. ⁴⁵, the peak of the CIB-induced excess is affected by the K-correction and the wavelength that was used to measure dust emission.

In stellar reddening based dust maps, the redshift distribution of the angular cross correlation signal is less straightforward and would depend on what classes of extragalactic objects and at what red-

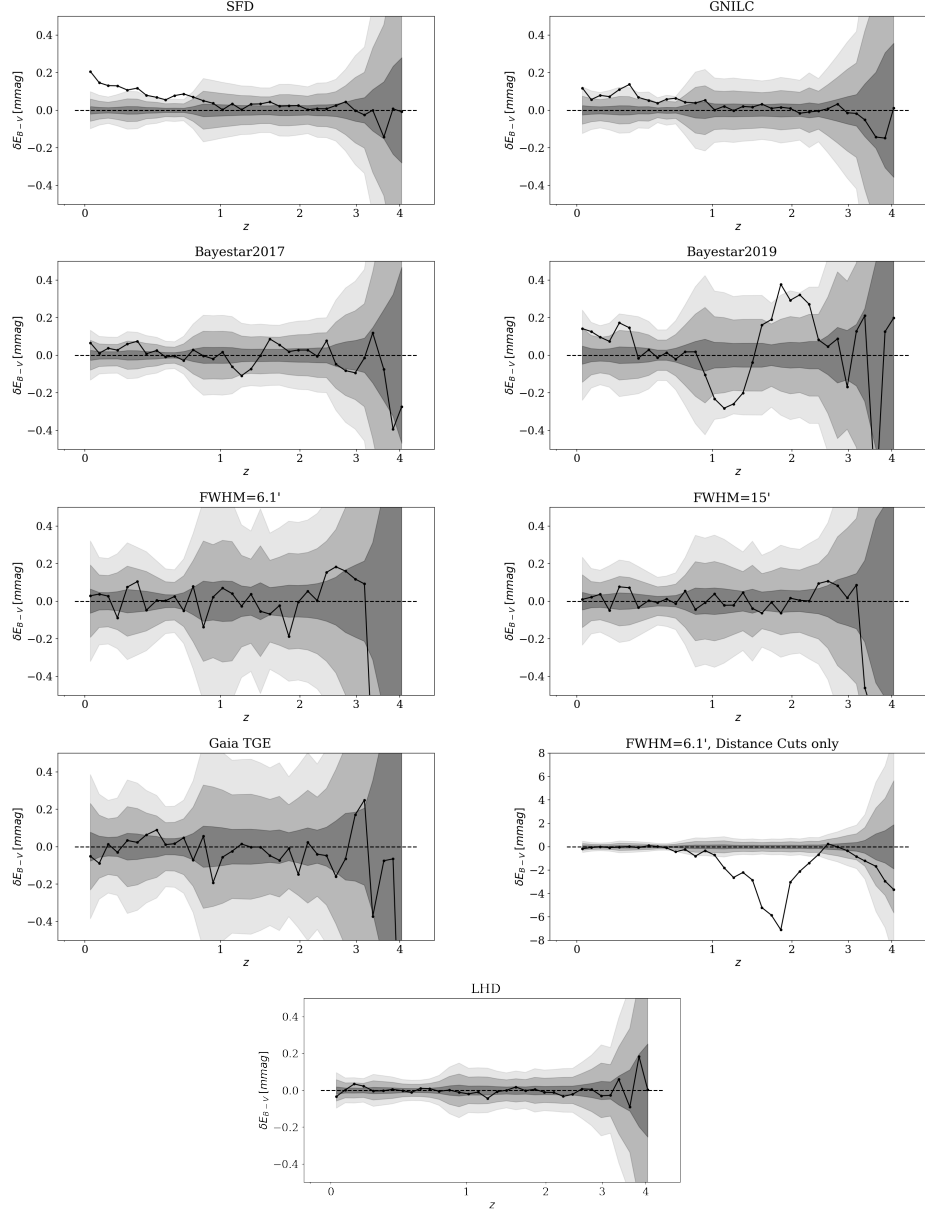


Figure 1.5: Excess extinction in maps at the positions of reference extragalactic objects in redshift bins $z = 0 - 4$ for emission-based maps (first row: SFD and the Planck 2016 GNILC dust emission map, **last row: the Lenz-Hensley-Doré map**), the **Bayestar** stellar-reddening based maps (second row: Bayestar17 and Bayestar19), our maps with FWHMs of 6.1' and 15' (third row). In the fourth row, we plot the signals for the Gaia TGE map and for the FWHM=6.1' map with only distance cuts applied. Without cuts intended to eliminate extragalactic objects, the contamination from QSOs exceeds 5σ for $1 < z < 2.5$. The sigma contours are derived by rotating the map by 100 angles in the range $30^\circ - 300^\circ$ and computing the root mean square error of the signals of the ensemble of rotated maps. For all maps, the signal was evaluated over the region $b > 50^\circ$.

shift have the greatest impact on biasing extinctions. Of the existing stellar-reddening based maps, Bayestar19, had significantly higher correlations than both Bayestar17 and the maps we construct. The factors that can potentially affect the level of correlation in the maps can include stellar selections performed on catalog objects, the information used to derive posterior extinction estimates as well as the reconstruction scheme. The discretization of Bayestar19 into units of 0.01 mag makes the map noisier, particularly at the higher latitudes where we evaluate the angular cross-correlation signal. This is imprinted in the larger noise envelope and possibly accounts for the excess correlated reddening in Bayestar19 relative to Bayestar17. The maps we construct (FWHM=6.1' and FWHM=15' in Figure 1.5) have correlations that are largely around the 1 sigma level, except in the highest redshift bins ($3 < z < 4$) where the signal is at the 3 sigma level and around $z \sim 2$ in the FWHM=6.1' map. Earlier stellar reddening based maps possessed negative correlations in the $z = 1.1 - 1.3$ range, and positive correlations at the $z = 1.6 - 2.2$ range at the ~ 5 sigma level in Bayestar19 and the $\sim 1 - 3$ sigma level in Bayestar17. The correlated reddening deficit at around $z \sim 1.2$ matches the increase in the QSO distribution observed in Figure 1.1. The Gaia TGE map also possesses correlations that are largely within the 1 sigma level, with a slightly larger dip near $z \sim 1$. The noisier star-based maps have larger error contours since the noise in the map propagates into the signal's estimator, as in Figure 1.4. It is also interesting that all maps appear to have underestimated signals at redshifts between $z = 3 - 4$, however, the sparsity of reference objects at higher redshifts makes the contribution of the noise high too. The signal and the error bars both rise significantly at redshifts close to $z = 3$ owing to the sparsity of the reference sample used to compute the signal at higher redshifts. In the right column of the fourth row of Figure 1.5, we plot the angu-

lar cross-correlation of a map at FWHM=6.1' without any of the extragalactic object-removal cuts applied and *only* the distance based cuts. There is a significant negative correlation peaking near a redshift of $z = 2$. After the cuts are applied, this $z = 2$ correlation is now at less than 3 sigma.

1.5.3 COMPARISONS TO SFD

In this section, we examine how consistent our maps are with SFD, as a function of different extinction regimes and latitude.

Extinction: We restrict ourselves to the footprint with valid predictions for our maps, and the Bayestar17 and '19 maps – the intersection of the Pan-STARRS1 footprint with regions with Galactic latitude $b > 20^\circ$. The error propagation scheme we follow in Section 1.4 implicitly assumes uncorrelated errors, and as a consequence underestimates the error of the resulting map, particularly at lower latitudes and for the FWHM=15' map, since as the number of stars used to predict the extinction value for a given pixel increases the error on the map is reduced. In addition to the reconstruction variance derived from Equation 1.8, we also report a reconstruction variance calibrated with respect to the uncertainty-normalized deviation between the map at FWHM=15' and the value of SFD at pixels at $b > 60^\circ$. We do so by adding in quadrature a $\sigma_{sys} = 0.01$ to the predicted reconstruction sigma. At this value of σ_{sys} , the distribution of $(Map@FWHM = 15' - SFD) / \sqrt{Var@FWHM = 15' + \sigma_{sys}^2}$ has a standard deviation of 1.03. If the values of the noise map were truly randomly distributed about the corresponding SFD values, we would have expected a standard deviation of 1. We only chose to set this additive error using higher latitudes where we expect the signal to be reasonably low. One would also not in general expect the uncertainty-

normalized bias to have a spread of 1 across the full sky and particularly at lower latitudes with higher extinction and more complex dust morphology where the map’s resolution would matter more. We examine the bias of pixels in the map, and the uncertainty-normalized bias $((Map - SFD)/\sigma(Map))$ of pixels in the map as a function of the extinction values in each pixel in SFD for the overall footprint, in Figure 1.6. The 2D histogram is normalized for each SFD extinction bin, and the values along the x axis for SFD span the 2nd to the 98th percentile of SFD values in our footprint. The values are most consistent at low values of SFD ($< 0.2\text{mag}$), which accounts for the majority of the pixels in the footprint we cover. At higher extinction values, all three stellar-reddening maps systematically underpredict extinction relative to SFD. We also note that while such comparisons are useful to gauge consistency across maps, SFD is not a ‘ground truth’ in this comparison. Rather, SFD is an emission-based map while the stellar-reddening based maps measure the effect of extinction — which is the effect that we actually want to map. Furthermore, Ref. ²⁴⁴ examined SDSS galaxy number counts at lower galactic latitudes, while Ref. ¹⁶ compared SFD extinction predictions to extinctions derived from the Taurus dark cloud complex using background stars’ color excesses and counts and ISSA 60 and 100 micron images. Both works found that SFD overpredicts extinction at $E_{B-V} > 0.15\text{mag}$. Other work has also found deviations in SFD: e.g. Ref. ¹⁶⁶ identified corrections to SFD by using quiescent galaxies as standard crayons (objects whose color is known) and found that SFD is underpredicted in regions at lower latitudes and $E_{B-V} < 0.15\text{mag}$.

Latitude: We examine how consistent our maps are with SFD, in $N_{\text{SIDE}}=32$ pixels at different latitudes in the Galactic Northern Hemisphere. We query all $N_{\text{SIDE}}=2048$ pixels lying within $N_{\text{SIDE}}=32$ pixels closest to a given latitude for the following latitudes:

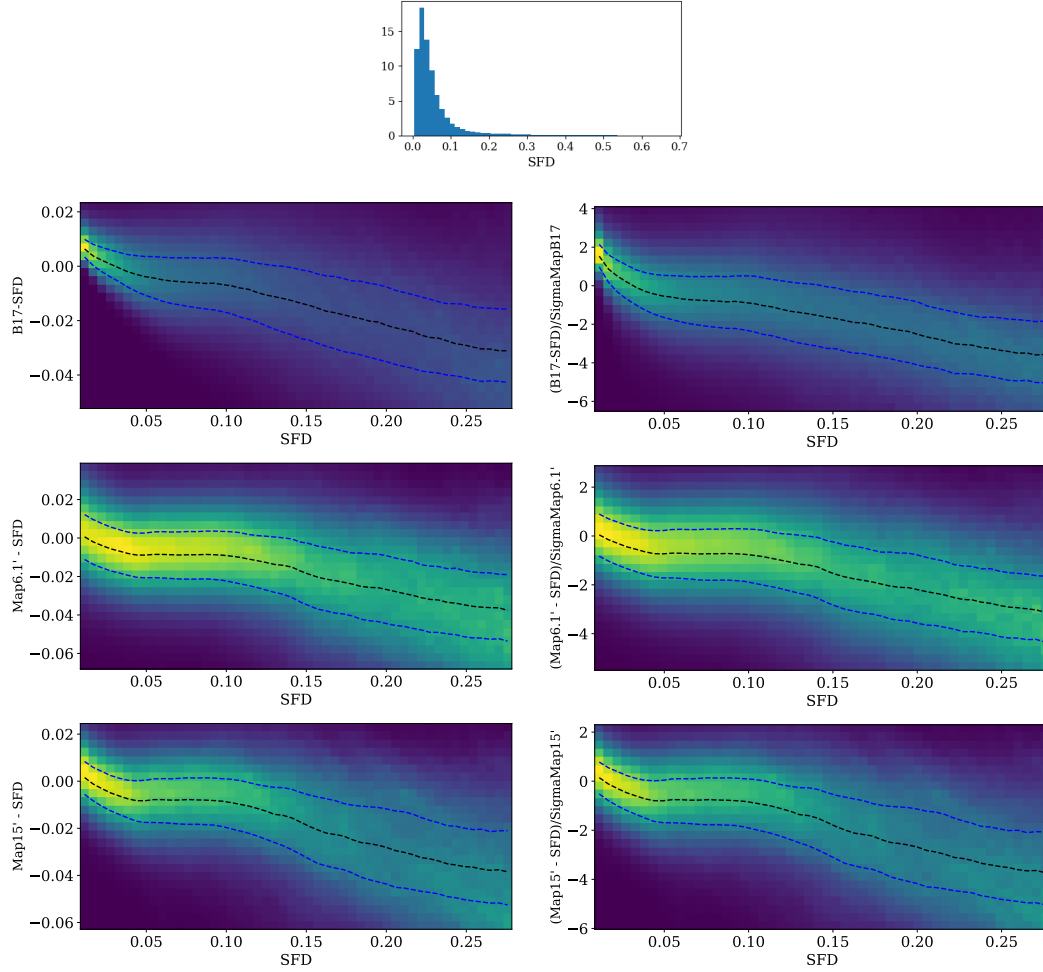


Figure 1.6: Histogram of SFD extinction values (top row). Two-dimensional histograms of the bias (left column) and uncertainty-normalized bias (right column) for Bayestar17, and our maps at FWHM=6.1' and FWHM=15' in the top, middle and bottom rows respectively. The zero point and uncertainty corrections have been applied here. The histograms are normalized such that the pixel values along the y axis for each value on the x axis (corresponding to a single SFD extinction bin) sum to 1. The dotted lines denote the 25th, 75th (blue) and 50th (black) bias percentiles for each SFD extinction bin.

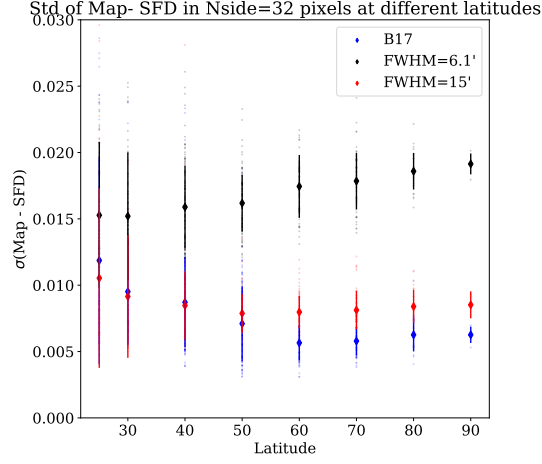


Figure 1.7: Standard deviations of the offsets of the map relative to SFD for all testbeds. Each datapoint is the standard deviation over the difference between the input map’s values and SFD’s values in $N_{\text{side}}=2048$ pixels contained within a single $N_{\text{side}}=32$ pixel.

$b = [25^\circ, 30^\circ, 40^\circ, 50^\circ, 60^\circ, 70^\circ, 80^\circ, 90^\circ]$. For each of the $N_{\text{SIDE}}=32$ pixels in each latitude band,

we examine the standard deviation of the difference of the map’s extinctions and SFD extinctions.

The mean and standard deviation of these values is visualized in Figure 1.7. As we move from lower to higher latitudes, with lower dust and relatively fewer stars, the $\text{FWHM}=15'$ map is less noisy than the one with $\text{FWHM}=6.1'$. In this limit, the less smoothed map is approximately 2.5 times noisier than the map with more smoothing, as would be expected from the ratio of the smoothing scales.

1.6 DISCUSSION

Contamination of Galactic extinction maps by large scale structure could potentially impact a range of cosmological measurements. Ref.⁴⁵ demonstrate how galactic dust map cross-correlations with LSS could bias lensing-induced correlations between foreground lenses and background objects and parameter constraints derived from luminosity distances. In the latter case, the bias is still mild, and

at the 0.5% level for Ω_m and w .

Ref.¹¹⁵ found that the density of DESI Emission Line Galaxy (ELG) samples and Quasi-Stellar Object (QSO) samples decreases and increases respectively with extinction at a level of nearly 10% in some extinction bins. In addition to the correlations we focus on in this paper, variation in Galactic extinction and stellar density can also affect the apparent distribution of extragalactic objects and add power on larger scales that are particularly important for constraints on the primordial non-Gaussianity parameter f_{NL} ^{190,185}. Cross correlations, such as those between CMB lensing maps and LSS tracers such as galaxies could also potentially be affected since Galactic dust serves as a common foreground for the CMB, in terms of its emission and for the magnitudes of galaxy catalog objects, in terms of its extinction. Ref.⁴² found that the galaxy-CMB lensing cross power spectrum can be biased up to a few percent by extinction even when the bias on the galaxy autopowerspectrum is at the sub percent level.

The maps we release could serve as a systematics cross-check in several ways. It would be of interest, for example, to examine whether different choices of maps, including the maps we construct, affect the magnitude and direction of these trends.

This work can be built upon in several ways. One of the key limitations of stellar-reddening based maps is their higher noise. In terms of the statistical reconstruction, better regularization could help reduce noise and enable us to derive better posterior uncertainties. The advent of data from upcoming surveys such as LSST would increase the number of stars observed by an order of magnitude and be able to observe much fainter stars ($r \sim 27$)¹⁰². LSST data might also render other novel and complimentary ways of constraining galactic dust extinction possible. Ref.³³ ex-

explored the possibility of simultaneously deriving Galactic dust extinction maps at resolutions of $7'$ and 1° as well as large scale structure overdensities with a simulated LSST galaxy catalog, a task hitherto considered impossible at resolutions comparable to that of existing emission or extinction-based dust maps because of limitations in depth or sky coverage. With their approach, correlated noise, including correlations between LSS and extinction, become dominant at resolutions of $7'$. They further outline a Bayesian scheme to combine existing dust maps as a prior and galaxy properties as a likelihood to derive a joint posterior distribution over LSS and Galactic extinction. It would be interesting to examine the possibility of combining maps with lower correlations with extragalactic structure such as the ones we sought to make in this paper, with constraints from galaxy properties, to both combine the constraining power of stars and galaxies on the dust distribution at higher latitudes, and minimize correlations with LSS. Better star-galaxy-QSO separation approaches constitute the common underpinning for both the approach we take and the one in Ref. ³³, a task that would become more challenging as we go to fainter magnitudes.

While we were preparing this draft, Ref. ⁶³ released a total galactic extinction map using Gaia DR3 sources classified as stars from the Discrete Source Classifier (DSC) module. Our work differs from that in several regards: the information used to exclude extragalactic objects, reconstruction choices – their maps have varying HEALPix resolution, ranging between HEALPix levels 6 and 9, the portion of the sky covered – they cover the sky at $|b| > 5^\circ$ and the stellar inference framework used – they use Ref. ¹¹ to derive extinctions to each source. Figure 1.8 plots SFD, Bayestar17, the Gaia total galactic extinction map, and our maps at FWHM= $6.1'$ and $15'$ over six different patches of the sky. To derive extinction estimates for the Gaia TGE map we divide the map, which reports

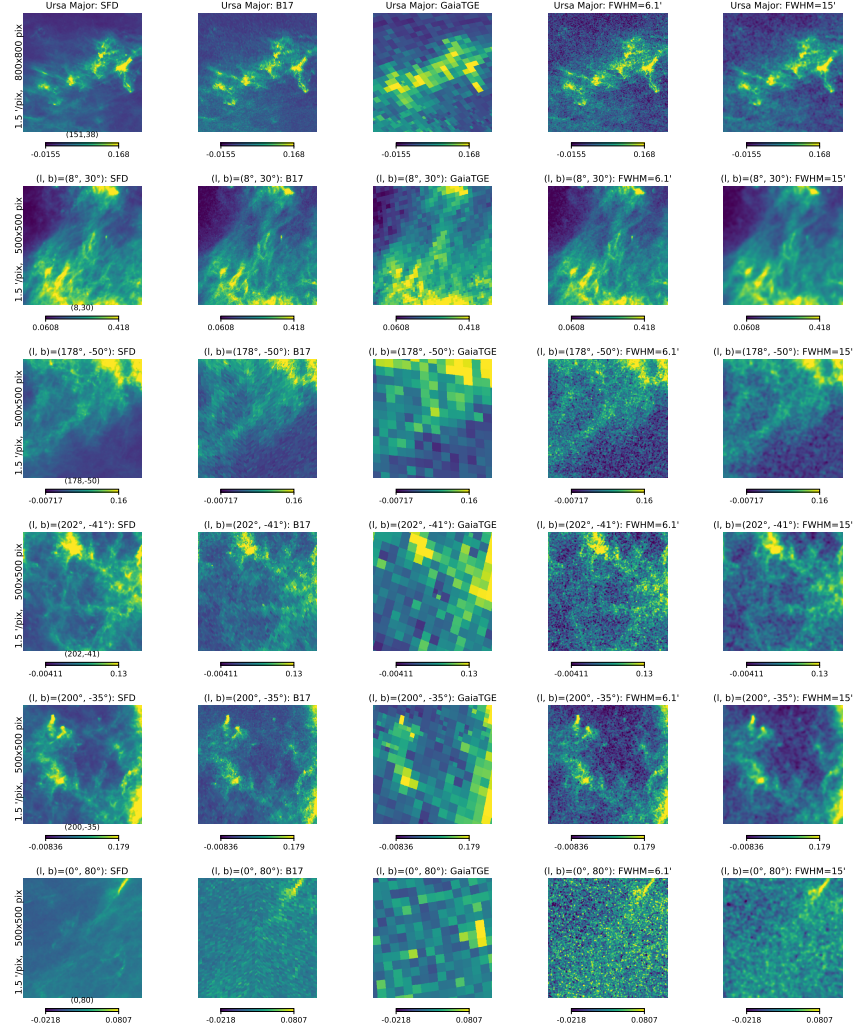


Figure 1.8: Comparison of SFD, Bayestar17, the Gaia TGE map and the reconstructions at FWHM=6.1' and FWHM=15' over different patches of the sky, with the colorscale set by the 1st and 99th percentile of the pixel values of the FWHM=6.1' map for the area under focus.

extinction at 541.4 nm assuming the Fitzpatrick extinction law with $R_V = 3.1$. We further multiply the map by a factor of 0.86 to account for the multiplicative factor that SFD has been multiplied by, since in Ref. ⁶³ a linear fit between TGE A_0 and SFD A_V had a slope of 0.98. Having multiple extinction maps allows us to test for sensitivity to differences and assumptions of any single stellar inference framework or source selection scheme.

1.7 CONCLUSION

Minimizing systematics in cosmological analyses is important to be able to harness the full power of datasets from DESI and LSST in the near future. To this end, we focused on examining the question of whether removing suspected extragalactic objects from stellar catalogs reduces correlations with large scale structure, and release two resulting Galactic extinction maps. We cross-match to SDSS DR17 data and learn the boundaries of regions populated by stars in PS1 *riz* - WISE W1 and SDSS *ugr* space. We additionally have cuts on Gaia EDR3 parallax and the reduced chi-squared statistic of the objects in the catalog with stellar models. Since we choose to reconstruct our maps in two dimensions, and wish to use only stars lying behind the sheet of integrated Galactic extinction over the sky, we add distance cuts to select only stars that we expect to lie behind dust in the Galaxy. This set of stellar selections reduces QSO contamination from 15% to $\sim 0.1\%$ on a test set of spectroscopically matched objects in the Southern Galactic Hemisphere. We then apply these cuts on the set of stellar catalog objects and construct two dust maps at a Full-Width Half-Maximum of 6.1' and 15' for the intersection of the PS1 footprint and the region with $|b| > 20^\circ$. We evaluate the an-

gular cross-correlation of these maps and for existing dust maps, using the clustering-redshift based technique, and find that our map has lower correlations than most existing stellar-reddening and emission-based maps.

1.8 CODE AND DATA AVAILABILITY

We make the code used for this work available on Github at **HighLatMaps**. The maps used for this work can be found here at this Zenodo link: <https://doi.org/10.5281/zenodo.7411344>

1.9 ACKNOWLEDGEMENTS

We are especially grateful to Gregory M. Green and Catherine Zucker for help with running the Bayestar stellar inference code and useful discussions. We thank Tanveer Karim, Andrew K. Saydjari, Joshua S. Speagle, Justina R. Yang and Ioana Zelko, for helpful discussions and inputs. We thank Yi-Kuan Chiang, Arjun Dey, Daniel Eisenstein, Ashley Ross, Eli Rykoff, David Schlegel and Martin White for insightful conversations on this work. This work was supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions). D.P.F. acknowledges support by NASA ADAP grant 80NSSC21K0634 “Knitting Together the Milky Way: An Integrated Model of the Galaxy’s Stars, Gas, and Dust.” C.F.P. acknowledges the support of NIH R01.

The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS

Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/-California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (<https://www.cosmos.esa.int/gaia>), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions.

SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration.

Software: GNU Parallel²¹⁹, NUMPY⁸⁹, ASTROPY¹⁷, HEALPIX²⁴⁷, SCIKIT-LEARN¹⁶⁵, PANDAS¹⁶⁰, MATPLOTLIB¹⁰⁰, DUSTMAPS⁸⁰

The following chapter and associated appendix originally appeared as Mudur, N. and Finkbeiner, D.P., 2022. Can denoising diffusion probabilistic models generate realistic astrophysical fields? arXiv: 2211.12444 *Machine Learning and the Physical Sciences workshop, NeurIPS 2022* DOI: <https://doi.org/10.48550/arXiv.2211.12444> (CC BY 4.0) *The title was updated.*

Diffusion Models for Astrophysical Fields

Score-based generative models have emerged as alternatives to generative adversarial networks (GANs) and normalizing flows for tasks involving learning and sampling from complex image distributions. In this work we investigate the ability of these models to generate fields in two astrophysical contexts: dark matter mass density fields from cosmological simulations and images of interstellar dust. We examine the fidelity of the sampled cosmological fields relative to the true fields using three different metrics, and identify potential issues to address. We demonstrate a proof-of-concept application of the model trained on dust in denoising dust images. To our knowledge, this is the first application of this class of models to the interstellar medium.

2.1 INTRODUCTION

Generative models of astrophysical and cosmological fields can serve a multitude of purposes. Cosmological simulations take several thousand CPU hours to run and can only be generated for a limited set of parameters. There thus exists a demand for emulators: frameworks that can generate summary statistics⁹² or the underlying cosmological fields^{71,105} conditional on an input cosmological / astrophysical parameter set. This can accelerate parameter inference approaches that require evaluating likelihoods at intermediate values. In the context of observed astrophysical fields, statistical descriptions capable of capturing the non-Gaussian nature of the interstellar medium would aid component separation problems encountered in searches for the B -mode of the Cosmic Microwave Background¹⁸¹ and statistical regularization for interstellar dust mapping^{82,123}.

Score-based generative models²¹⁵ have witnessed a surge in interest because of findings that show that they surpass GANs in terms of image fidelity⁶⁵ and their ability to produce realistic images conditional on text inputs¹⁹⁶. These models learn the gradient of the probability density of the data to learn a generative model of the data distribution. A subset of this class includes denoising diffusion probabilistic models (DDPMs)⁹⁷. Ref.²¹² used DDPMs to generate galaxy images. Ref.^{183,182} used the denoising score matching framework to learn the non-Gaussian component of a prior on weak lensing convergence maps from simulations.

In this work we investigate two applications of DDPMs – one to simulation data products and one to images of interstellar dust. We train models to generate dark matter density fields from a simulation suite on grids of 64x64 and 128x128 pixels. We then compare five summary statistics be-

tween samples from the trained models and the real simulation fields. While we currently generate these fields unconditionally, this benchmarking is an important step toward using these models as emulators, and generating fields conditional on an input parameter vector. We then turn our attention to the real sky, and train a model to generate square patches from an interstellar dust map. We use the trained model as a denoising model and examine how well it can reconstruct the underlying image given a noisy input.

2.2 DDPM BACKGROUND

In this section, we briefly review the denoising diffusion probabilistic model formulation in Ref.⁹⁷.

A DDPM consists of a forward and a reverse diffusion process, over a fixed number of time steps,

T , where $\mathbf{x}_0$ is a draw from the image distribution and $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. The forward diffusion pro-

cess is defined according to a variance schedule $\{\beta_t\}$. Thus $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$

and $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t = \prod_{t'=1}^t 1 - \beta_{t'}$. The neural network $\theta(\mathbf{x}_t, t)$

parameterizes the reverse diffusion process: $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$ where $\mu_\theta(\mathbf{x}_t, t) =$

$\frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\theta(\mathbf{x}_t, t))$ and $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. A simplified loss function is minimized where

$L_{t-1} = ||-\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}, t)||^2$ with $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. Thus for each batch a set of timesteps t is

uniformly sampled from $t \sim U[1...T]$ to minimize L_{t-1} .

2.3 GENERATIVE MODELS FOR DARK MATTER DENSITY FIELDS

DATASET The CAMELS Multifield Dataset (CMD)²³⁰ from the Cosmology and Astrophysics with Machine-Learning Simulations (CAMELS) dataset²²⁹ was used for this application. The CMD includes an ensemble of thirteen two-dimensional physical fields for 1000 different simulation parameter vectors where each vector consists of 2 cosmological and 4 astrophysical feedback parameters. Every unique parameter vector has 15 samples and the full data set thus consists of 15000 samples of each physical field. We work with the log (base 10) of the cold dark matter mass density field from the IllustrisTNG hydrodynamical simulation at $z = 0$, at two grid sizes (64x64 and 128x128) binned down from the original 256x256. Each side of an image corresponds to $25h^{-1}$ Mpc. For the 64x64 model, we use fields corresponding to the first 60% of the parameters as our training data. We augment our fields with rotations and flips, to yield 54000 (9000x6) train fields. For the 128x128 model, we use fields corresponding to the first 70% of the parameters as our training data. We augment our data with rotations, flips and translational shifts (since the data has periodic boundary conditions). We thus have 252000 (10500x24) train fields. We apply a minmax transform that scales the minimum and the maximum pixel intensity of the full training set to $[-1, 1]$.

TRAINING DETAILS We train two models, one at each resolution, for 60k iterations (batch updates). In both cases, we use a forward diffusion process parameterized by a linear variance schedule lying in the range $[10^{-4}, 2 \times 10^{-2}]$, $T=2000$, a batch size of 40 images, and the Huber loss

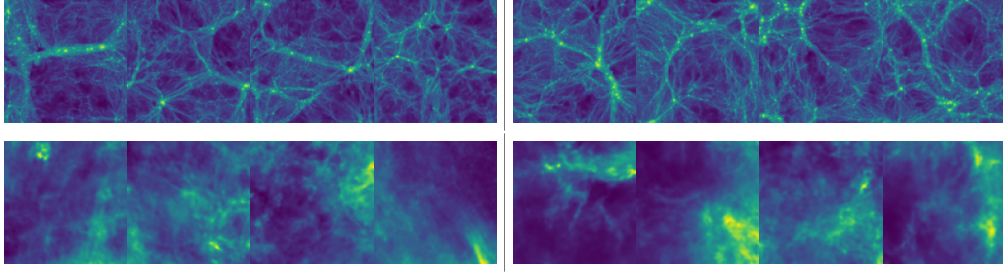


Figure 2.1: Four log cold dark matter mass density fields from the training data (top left) and from the sampled model (top right) at 128×128 . Four samples of dust from the training data (bottom left) and from the trained model (bottom right).

in place of the L2 loss with the Adam optimizer¹¹³. We used a learning rate of 5×10^{-4} for the model at 64×64 and 2×10^{-4} for the model at 128×128 and saved checkpoints every 2000 iterations, to enable sampling from multiple models. For the 64×64 case, we train models with 3 different seeds. We use code blocks and architecture from Hugging Face’s The Annotated Diffusion model¹⁸⁷, <https://github.com/lucidrains/denoising-diffusion-pytorch> and Ref.⁹⁷. The architecture we use is similar to that in Ref.⁹⁷, and consists of a U-Net¹⁸⁹ with 4 down and up-sampling blocks consisting of 2 ResNet blocks²⁴⁵, group-normalization²³⁸, and attention^{227,207}. We use the Weights and Biases framework²⁶ for our experiments. The code for experiments in this paper is available at the following repository: <https://github.com/nmudur/diffusion-models-astrophysical-fields-mlps>.

SUMMARY STATISTICS Four samples from the trained model at 128×128 are plotted in the top right panel of Figure 2.1. We consider three summary statistics – the power spectrum, the normalized pixel intensity histograms and the three Minkowski functionals in Figure 2.2 to gauge consistency between the sampled fields and the true image distribution. Minkowski functionals^{203,202}

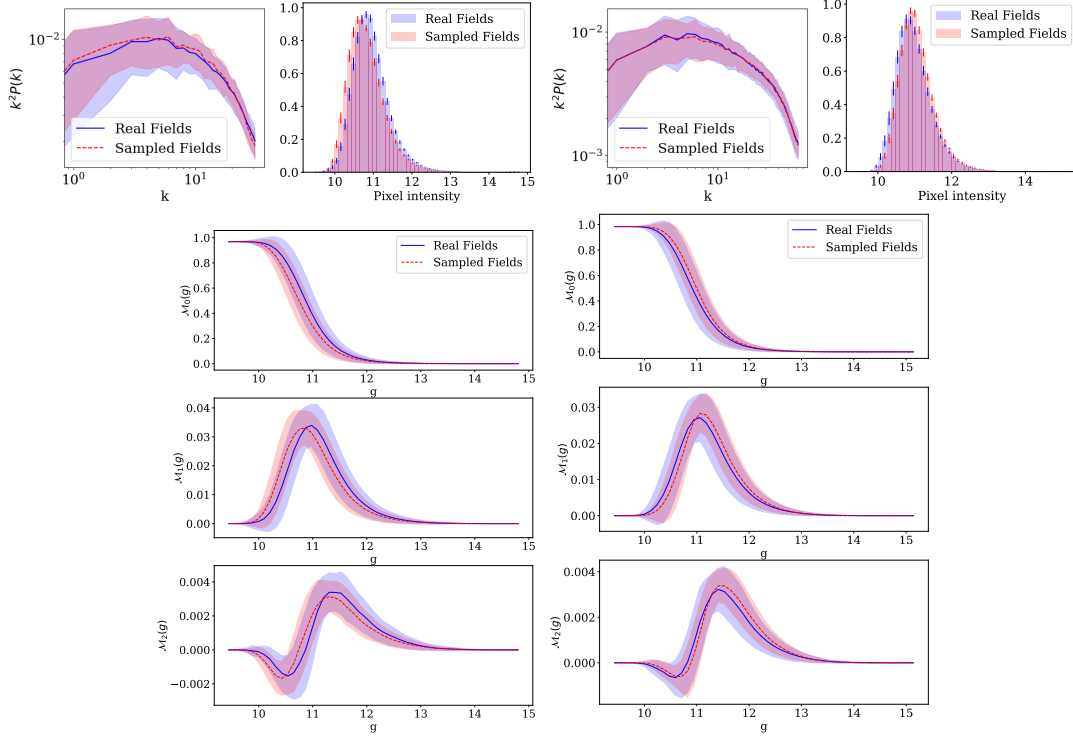


Figure 2.2: Power spectra, normalized pixel histograms and Minkowski functionals for 100 draws from the real fields and the trained models at 64x64 (left) and 128x128 (right). The envelopes in the power spectra and the Minkowski functionals' panels represent the standard deviation of the value of the statistic in each bin. The height of the bars in the pixel histogram is the mean height of the histogram bin and the error bar is the standard error over all 100 samples.

are topological descriptors of fields sensitive to correlation functions beyond the second order and can be used as metrics to gauge how similar the statistics of the generated fields are to those of fields from the true distribution e.g.,^{218,177}. They are computed as integrals over excursion sets with pixels whose intensity is greater than a value g . In 2D, $\mathcal{M}_0$, $\mathcal{M}_1$, $\mathcal{M}_2$ reflect the number of pixels in the excursion set (area), the length of its boundary (perimeter) and the number of holes. We use QuantImPy³⁰ to compute these functionals.

The loss function used to train these models does not directly enforce the summary statistics of

the generated images to exactly match the summary statistics of the training distribution. Thus, while the mean of the statistics of the generated samples across checkpoints typically lie within the one standard deviation envelope of the power spectra and Minkowski functionals of the true distribution, convergence in terms of the loss does not imply that the converged models are stationary with respect to the distribution of the summary statistics of the generated images. We thus draw 100 samples from each of the last 10 checkpoints of the trained diffusion models and from the real (train) dataset. Since the variability across checkpoints is similar across all three seeds for the 64x64 models, we train a single model for the 128x128 case and sample from its last 10 checkpoints. For the samples across all 10 checkpoints, the mean absolute fractional difference in the power spectrum is typically around 10% (std. dev. 5% for 64x64) and 15% (std. dev. 10% for 128x128) and is higher at the lowest and highest k -bins, which are most affected by cosmic variance and noise, respectively. For both the 128x128 model and all 3 runs of the 64x64 model, at least one checkpoint with a mean (over all bins) absolute fractional difference of less than 5% could be found. As in Ref. ¹⁴⁶ and Ref. ²¹⁸ that use adherence to summary statistics as a model selection criterion, we identify the checkpoints with the lowest absolute fractional error in the power spectrum and the best agreement with the Minkowski functionals. The statistics plotted in Figure 2.2 correspond to these models. For the 64x64 case, we plot the statistics for the worst of the three best-case models (one for each seed). While the mean and the spread of the power spectra appear to be largely consistent for these models, the Minkowski functionals are more sensitive to differences between the true and the generated samples. We intend to explore whether these differences and the lack of convergence to a distribution stable with respect to the summary statistics can be mitigated with different design choices,

or whether more fundamental changes to the method are required.

2.4 A GENERATIVE MODEL FOR INTERSTELLAR MEDIUM FIELDS

DATASET AND TRAINING DETAILS The dataset consists of 12482 images from the Schlegel-Finkbeiner-Davis (SFD)²⁰¹ map of interstellar dust extinction inferred from emission at 100 microns. We restrict ourselves to images with extinction $E_{SFD} < 3$. Each image is 128x128 and spans an area of $(6.4^\circ)^2$ on the sky. The validation set consists of images lying in Galactic longitude $0^\circ < l < 42^\circ$ and the test set (held out for future validation) consists of images lying between $200^\circ < l < 240^\circ$. All other images (roughly 69%) belong to the train distribution. The train images are augmented with rotations and flips, yielding 68048 images (8506x8). We apply a minmax transform that scales the minimum and the maximum pixel intensity of *each image* to $[-1, 1]$. We peg the transform to each image because images of interstellar dust span a larger dynamic range and pixel intensities in two images are much more likely to be dissimilar than for the cosmic web. We train our models for 42k iterations with a learning rate of 6×10^{-5} . All other training details are the same as in Section 2.3. Four sampled images are plotted in the bottom right panel of Figure 2.1. For both the generated cosmic web images and the dust images we see that the models are able to capture a rich variety of structure.

DENOISING TESTS We select a filamentary field from the validation dataset, whose standard deviation corresponds to roughly the 50th percentile of the standard deviation across all images in the validation set. We add $\mathcal{N}(0, \sigma^2)$ noise such that σ is 20% of the mean of the intensity in the image.

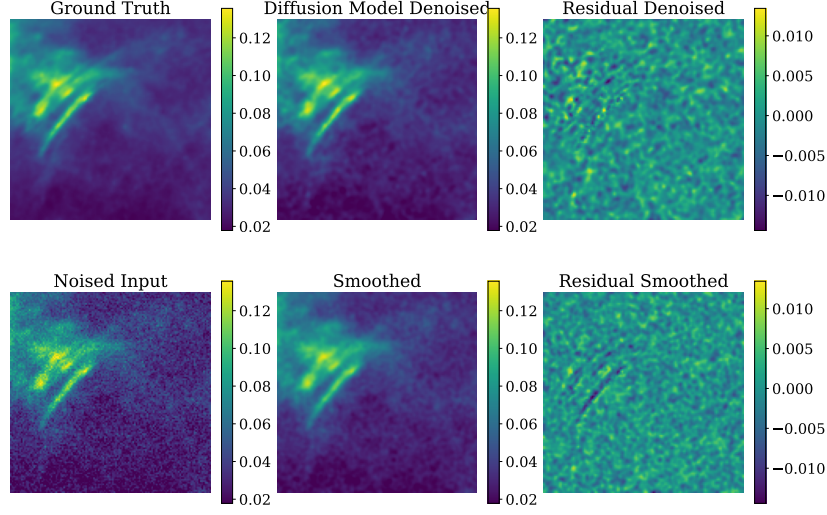


Figure 2.3: Ground truth, noised input, denoised image and its residual (denoised - truth), smoothed baseline and its residual (smoothed - truth).

The noisy input is scaled to $[-1, 1]$ and the timestep at which $\sqrt{1 - \bar{\alpha}_t}$ is closest to the corresponding scaled sigma σ_{tr} is identified for each image. We then iteratively sample from $p_\theta(x_{t-1}|x_t)$ from $t = t_{\sigma_{tr}}$ to $t = 0$ to derive the denoised image. As a baseline, we find the corresponding Gaussian filter that would reduce the RMSE in the low-signal portion on the bottom right of the image by the same factor (3.3). Figure 2.3 plots the denoised images with the model and the baseline. The correlation of the residual with filaments is significantly lower with the diffusion model than with the baseline.

2.5 CONCLUSIONS AND FUTURE WORK

In this work, we investigated whether DDPMs are able to learn and sample from the image distribution for two astrophysical fields — one from simulations and the other from real interstellar dust maps — in a ‘physically meaningful’ way, gauged by two different yardsticks. In the case of mod-

els trained to generate images of interstellar dust, the models are able to denoise highly filamentary images of interstellar dust while recovering underlying structure better than a smoothing baseline. Approaches such as Ref. ¹⁷⁷ involve learning generative models of interstellar dust using the wavelet statistics of a single image, whereas our description learns a prior using multiple images. The former approach can be useful in cases where limited training data is available while the latter approach is more likely to account for diversity and multiple modes of the image distribution. In the context of dark matter density fields from simulations, we examined three sets of summary statistics of cosmological significance. The ability to generate images with the same statistics is an important first step toward deploying these models as emulators. While DDPMs show promise in terms of our ability to find models that generate samples that are consistent up to 10% in the power spectrum, we intend to work on finding architectures or models with inductive biases that prioritize convergence to distributions that are stationary with respect to these summary statistics.

2.6 BROADER IMPACT

Several papers have examined and improved the performance of score-based generative models for standard machine learning datasets using standardized metrics such as the Fréchet Inception Distance ⁹⁶, demonstrated their ability to generate photo-realistic images conditional on text inputs ¹⁷⁶ and identified issues to work on, such as their slower sampling speed, relative to GANs. This raises the question – how can these models help accelerate science? As a first demonstration of the use of these models to generate fields from the interstellar medium, we hope that this paper can motivate

both potential use cases for these models in astrophysics and cosmology as well as work on imbuing these models with physical inductive biases that make them more suitable for applications where practitioners care about recovering specific summary statistics. While our focus here has been on astrophysics, the questions we consider and the desire for high-fidelity generative models capable of sampling distributions of images faster than simulations have wider applications across the physical sciences¹¹¹ — from the geophysical sciences to high energy physics¹⁵⁷.

2.7 ACKNOWLEDGEMENTS

We thank Shuchin Aeron, Carolina Cuesta-Lazaro, Tanveer Karim, Andrew K. Saydjari, and Justina R. Yang for helpful discussions. This work was supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions).

The following chapter and associated appendix was originally published as Mudur, N., Cuesta-Lazaro, C. and Finkbeiner, D.P., 2024. Diffusion-HMC: Parameter Inference with Diffusion-model-driven Hamiltonian Monte Carlo. *The Astrophysical Journal*, 978(1), p.64. DOI: 10.3847/1538-4357/ad8bc3 (CC BY 4.0). *Some typographical errors were fixed and a more detailed explanation of an equation in the appendix added. The placement of some figures / tables and their referencing text was updated.*

3

Diffusion-HMC: Parameter Inference with Diffusion-model-driven Hamiltonian Monte Carlo

DIFFUSION GENERATIVE MODELS HAVE EXCELLED at diverse image generation and reconstruction tasks across fields. A less explored avenue is their application to discriminative tasks involving regression or classification problems. The cornerstone of modern cosmology is the ability to generate predictions for observed astrophysical fields from theory and constrain physical models from

observations using these predictions. This work uses a single diffusion generative model to address these interlinked objectives – as a surrogate model or emulator for cold dark matter density fields conditional on input cosmological parameters, and as a parameter inference model that solves the inverse problem of constraining the cosmological parameters of an input field. The model is able to emulate fields with summary statistics consistent with those of the simulated target distribution. We then leverage the approximate likelihood of the diffusion generative model to derive tight constraints on cosmology by using the Hamiltonian Monte Carlo method to sample the posterior on cosmological parameters for a given test image. Finally, we demonstrate that this parameter inference approach is more robust to small perturbations of noise to the field than baseline parameter inference networks.

3.1 INTRODUCTION

Ongoing and upcoming missions, such as the Dark Energy Spectroscopic Instrument [DESI](#), the Vera C. Rubin Observatory’s [Legacy Survey of Space and Time](#) and the [Nancy Grace Roman Space Telescope](#) will map the cosmos at unprecedented resolution and volume. This has created a proportionate demand for simulations that can generate predictions from theory. Cosmological simulations, however, are expensive to run, and can only be generated for a limited set of initial conditions and points in parameter space.

The canonical summary statistic used for parameter inference is the two point correlation function, or the power spectrum Pk at large, linear scales where perturbation theory holds. At smaller

scales, however, gravitational collapse induces non-Gaussianity in the fields. This means the information content of cosmological fields is not fully captured by the power spectrum at the large scales alone. For example, recent work^{87,150} derived much stronger constraints on σ_8 by going beyond the two point correlation function at linear scales and analyzing non-linear modes at smaller scales ($k > 0.25 \text{ hMpc}^{-1}$) or by extracting information at the field-level. A multitude of other statistics – such as the marked power spectrum, the bispectrum, the wavelet scattering transform, and void probability functions – have also been devised^{225,158,177,88} in an effort to capture higher order correlations in non-Gaussian fields. Previous work^{92,146,158,226,206} has addressed the prohibitive cost of simulations by creating emulators or surrogate models that learn to interpolate between predictions of a *specific* summary statistic between training points using formalisms such as Gaussian processes. More recently, simulation-based inference at the field level has also been used, where a likelihood model parameterized by a neural network is learned on the fields^{61,56}.

Generative models are a class of machine learning approaches that enable one to simulate the ability to draw samples from a complicated target probability density and include variational autoencoders, normalizing flows and generative adversarial networks (GANs). Diffusion generative models^{213,216} involve a forward diffusion (noising) process that transform samples from the target distribution to those from the standard normal. In the denoising diffusion probabilistic model (DDPM)⁹⁷, the noising process consists of a variance schedule β_t over a fixed number of time steps, T , that determines the incremental noise added to the image. Since the diffusion process can be formulated as a stochastic differential equation (SDE) the DDPM variance schedule corresponds to the discretization of this SDE. In the generative direction, a neural network or score model is used to

parameterize the reverse transformation.

Diffusion models are alternatively referred to as score-based generative models since parameterizing the reverse diffusion process is equivalent to learning the ‘score’ or $\vec{\nabla} \log p_t(x)$ of the data^{10,216}. Since in high dimensions, the target distribution invariably lies on a thin manifold, the incremental addition of the random noise, blurs the distribution and makes the score progressively easier to learn. Diffusion models are the underlying mathematical framework that have given rise to the photorealistic image generation successes of DALL·E and Stable Diffusion¹⁸⁸ and have been shown to mitigate mode collapse, a phenomenon, often encountered with GANs a generative model fails to generate multiple modes in a distribution. In scientific applications, they have been used for problems involving protein folding and ligand prediction and medical imaging reconstruction^{54,214}. They have been applied in astrophysics to reconstruction problems involving dust^{95,144}, cosmological simulations and initial conditions reconstruction^{122,191,153,56,143} and strong lensing problems^{104,182}.

In this work, we apply diffusion generative models to emulating cold dark matter density fields conditional on cosmological parameters and demonstrate that the trained model can also be used to derive tight and robust constraints on cosmological parameters. In Section 3.3, we examine the ability of the model to appropriately capture the statistics of the distribution of fields corresponding to different parameters, and further compare the effect of modulating a single parameter at a time on the statistics of the resulting fields in the true and the generated set. We then quantify the ability of the model to capture the full range of cosmic variance for a *single* parameter, as a means to assess the extent of mode collapse. In Section 3.4, we examine how the diffusion model’s approximate likelihood can be used to solve the inverse problem of constraining the cosmological parameters of a

given input field. We then use the Hamiltonian Monte Carlo^{69,148,25} method to draw samples from the estimated posterior on the cosmological parameters given an input field, and compare our estimates with a power spectrum baseline. A novel contribution of this work is our use of an HMC to sample a posterior consisting of an approximation to the diffusion model conditional likelihood to solve a downstream inference task. Finally, we demonstrate that the Diffusion-HMC based parameter inference estimates are more robust to perturbations composed of uncorrelated noise relative to the estimates from a discriminative neural network directly trained to estimate parameters.

3.2 DATASETS, ARCHITECTURE AND TRAINING

DATASETS We work with cold Dark Matter density fields at $z = 0$ from the IllustrisTNG^{149,168} suite from the CAMELS Multifield Dataset (CMD)^{230,229}. The diffusion model is trained to generate the minmax transform applied to the log (base 10) of these dark matter fields. The minmax transform is pegged to the minimum and maximum of the log of the entire dataset, $[9.42, 15.44]$. The dataset contains 1000 simulations for 1000 different cosmologies with 15 two-dimensional fields per simulation. The ‘cosmology’ is parameterized by a parameter vector with 2 cosmological (Ω_m and σ_8) and 4 astrophysical parameters. Each simulation tracked the evolution of 256^3 dark matter particles and 256^3 fluid elements and took around 6000 CPU hours to generate (see Ref.^{230,229} for more details). The fields span $25 \text{ Mpc}.h^{-1}$ on each side. We train on 70% of the parameters in the LH set, i.e. 700 parameters or 10,500 fields, and condition the model only on the cosmological parameters Ω_m and σ_8 . Since we use dark matter density fields in this study, we do not expect them to constrain

or contain much information about the astrophysical parameters. Example generated dark matter fields are shown in Figure 3.2.

DIFFUSION MODEL SETUP We follow the denoising diffusion probabilistic model (DDPM) formalism, in which a target image x_0 is transformed to a sample from $x_T \sim \mathcal{N}(0, \mathbb{I})$ over the course of $T = 1000$ timesteps. The forward diffusion process follows an incremental noise schedule $\{\beta_t\}$, and the noise is added in a variance-preserving way.

$$\begin{aligned} \text{For } t \in [0, T-1], q(x_{t+1}|x_t) &= \mathcal{N}(\sqrt{1-\beta_t}x_t, \beta_t\mathbb{I}) \\ \text{and } q(x_{t+1}|x_0) &= \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbb{I}) \\ \text{where } \bar{\alpha}_t &= \prod_{t'=0}^t 1-\beta_{t'} \end{aligned} \tag{3.1}$$

The score / noise-predictor model is a U-Net¹⁸⁹ similar to that used in Ref.⁹⁷. There are 4 down-sampling layers, with each layer consisting of 2 ResNet blocks²⁴⁵, group-normalization²³⁸, and attention^{227,207}. We use circular convolutions in the downsampling layers since the input fields have periodic boundary conditions. Each parameter is normalized to lie between $[0, 1]$ with respect to its range, $\Omega_m \in [0.1, 0.5]$, $\sigma_8 \in [0.6, 1.0]$. A multilayer perceptron (MLP) transforms the cosmology vector into a space with the same dimension as the time embedding, and each ResNet block additionally has an MLP conditional on cosmology. The variance schedule β_t is non-linear with smaller steps at smaller t and larger steps for larger values of t (see Appendix C.1). We train the model to generate the *log* of the fields, and randomly rotate and flip the image to account for these invariances.

During training, for each batch of images $\mathbf{x}_0$, a batch of timesteps is sampled uniformly along with a noise pattern $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. The loss function minimized is $\|\varphi(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, t, \theta)\|^2$ for each set of $\{\mathbf{x}_0, t, \theta, \varepsilon\}$, where θ is the parameter vector. Our implementation minimizes the Huber loss, which behaves as an L1 loss for values of the loss greater than 1 and a Mean Squared Error (MSE) loss otherwise. From the training curves, the loss during training is less than 1 throughout, so the loss being minimized is in effect the MSE loss. We used the Weights and Biases framework²⁶ to track experiments. The model has 31.2 million parameters.

TRAINING We first downsample the images by a factor of 4 and train the conditional diffusion model on these 64x64 images for 60000 iterations. Since the UNet is formulated in terms of relative downsampling, the same architecture can be applied to images with different resolution. To train the model to emulate 256x256 fields, we initialize the checkpoint with the weights of the 64x64-trained model after 60000 iterations and trained for over 400000 iterations. We found that initializing the 256x256 model with the weights of the 64x64 model and then training the model led to faster convergence.

The noise prediction loss does not fully capture sample quality and convergence, and we need an alternative metric to assess the quality of the generated samples Ref.²²⁰. We sampled 500 fields (with 50 fields for 10 different validation parameters) for the checkpoints after 200k, 220k, 240k, 260k, 280k, 300k, 320k and 340k iterations. Sampling a batch of 50 256×256 fields from our model takes 310 seconds (6.2s/field). We computed the reduced chi-squared statistic (Equation C.4) of the power spectrum of each generated field, s , with respect to the reference distribution comprised of

the 15 true fields for that parameter. We then compute the mean and standard error of these values across all parameters and sampled fields.

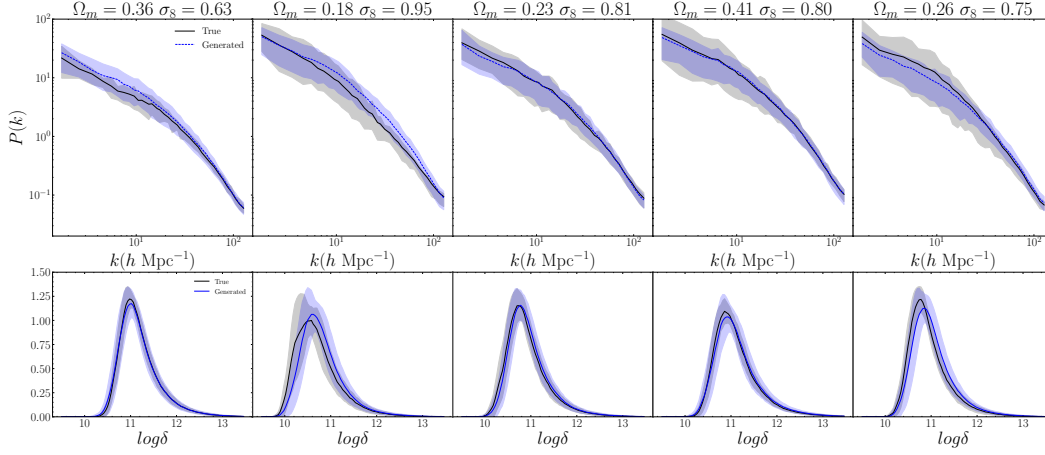


Figure 3.1: Generated fields at different cosmologies. **Upper Row:** Power spectrum of the unlogged fields for five different validation parameters. The lines depict the mean power spectrum and the envelope indicates the 16th and 84th percentiles of the distribution. True simulations are shown in black, generated in blue. **Lower Row:** Mean and standard deviation envelopes for the density histograms of the log fields.

While the diffusion generative model is trained to generate the fields in log space, the power spectra we compute here and in Figures 3.1- 3.3 correspond to the overdensity power spectra of the ‘linear’ ($10^{\text{GeneratedFields}}$) fields. The checkpoint corresponding to the 260kth iteration had the lowest value for the chi-squared statistic of the power spectra of the linear generated fields relative to the distribution of true fields, corresponding to 2.29 ± 0.49 (see Appendix C.2.2). To put this number in perspective, we can examine the effect of cosmic variance on this metric using a leave-one-out cross-validation approach, by computing the reduced chi-squared statistic of each sample of a *true* field, using the 14 other true fields corresponding to the same parameter as the reference distribution. The mean of this value across the 10 parameters is 1.70 ± 0.36 . We use the 260k checkpoint for our

analysis. We plot these values, along with the reduced chi-squared statistic of the power spectra of the *log* fields and the *p* values of the mean intensity in Appendix C.2.1.

3.3 SUMMARY STATISTICS

In this section we examine the consistency of our generated fields relative to the true fields using three sets of simulations under the IllustrisTNG CAMELS suite. The Latin Hypercube, **LH**, suite varies the largest range of cosmological and astrophysical parameters but has a limited number (15) of fields for each parameter, with all parameters varying randomly. The one-parameter, **1P**, suite consists of 15 fields with the same seed, and one dimensional variations of each parameter modulated systematically, while the others are fixed to the fiducial value. The Cosmic Variance, **CV**, set consists of 405 fields for the fiducial parameter value.

Varying cosmology in a Latin Hypercube (LH): We examine the consistency of the summary statistics of the distribution of true and generated fields for a given validation parameter from the LH set in Figure 3.1. We have 15 true fields and 50 generated fields for each parameter. We derived the boundaries of the envelope using the estimates of the 16th and 84th percentiles, while the solid line demarcates the mean of the distribution of power spectra in 35 log spaced *k* bins. The lower panel depicts the density histograms of the log fields. The envelopes are again derived using the percentiles, while the solid lines indicate the means of the histograms for the true and generated fields. The distribution of the power spectra and the density histograms of the true and the generated fields are in good agreement with each other.

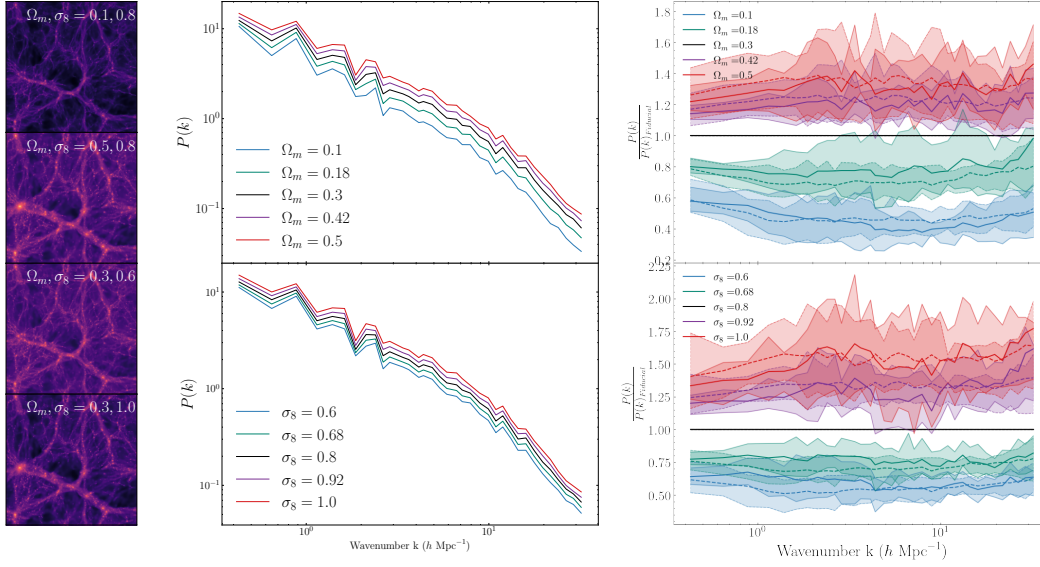


Figure 3.2: Generated ‘1P’ fields. **Left column:** Generated fields corresponding to the extreme values of each parameter for a single seed, with the other value held fixed at the fiducial value (0.3 for Ω_m and 0.8 for σ_8). **Middle column:** Power spectra of the generated fields for the same seed, for different values of each parameter, holding the other fixed. **Right column:** Mean and standard deviation for the ratio of the power spectra at the modified parameter value to the power spectra for the field at the fiducial parameter value (black) for 15 slices from the CAMELS dataset (solid) and 15 seeds for the generated fields from the diffusion model (dashed). The effect of modulating a parameter on the generated fields’ power spectra is consistent with that of the true fields.

Varying cosmological parameters one at a time (1P): In Figure 3.2, we generated ‘1P’ sets and examined whether the effect of modulating a single parameter, while keeping the others constant is the same as is observed in the 1P CAMELS suite. We sample 15 fields corresponding to 15 different seeds for each of the parameters. The fields corresponding to the same seed across parameters have the same position and orientation of their seeded structures as visible in Figure 3.2. For each seed, we then compute the ratio of the power spectrum of a field at a different parameter to the power spectrum of the fiducial parameter value, and compute the average and the percentile-based standard deviation envelopes across all seeds as depicted in the right column of Figure 3.2. The dashed (solid) line and envelope correspond to the generated (true) ratios. The ratios for the generated ‘1P’ set are in good agreement with those of the true ‘1P’ set, since the dashed lines are within the solid envelopes.

Reproducing cosmic variance (CV): The CV set has 405 (27×15) fields for the fiducial parameter value of $[0.3, 0.8]$ and varying initial conditions, designed to quantify the effect of cosmic variance. The CV set allows us to quantify the consistency between the second moments of the true and the generated distributions. In particular, it allows us to test the ability of the model to generate a diverse set of samples for the same cosmological parameters such that it reproduces the true underlying distribution at fixed cosmology.

We generate 405 samples from our trained diffusion model, compute their power spectra, and examine the standard deviation, and the correlation matrices of different k modes of the power spectrum in Figures 3.3 and 3.4. The correlation between the modes of the power spectra is largely consistent between the true and the generated samples, although the generated samples appear to

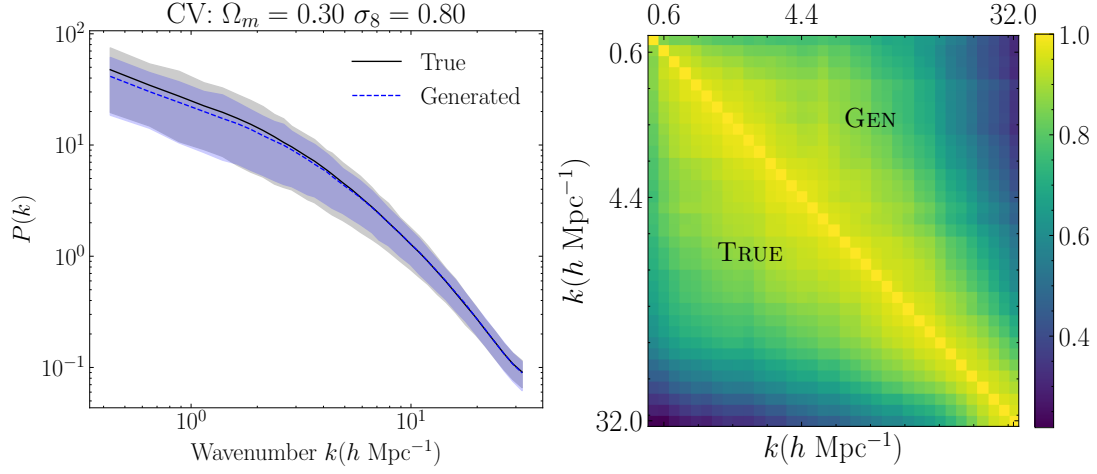


Figure 3.3: Generated ‘CV’ fields. **Left:** Power spectra of the 405 true and generated fields, with the mean, and 16 – 84th percentiles. **Right:** Correlation matrix of the power spectra of the true and generated fields. The lower triangular matrix corresponds to the correlation matrix of the true fields while the upper triangular matrix corresponds to the correlation matrix of the generated fields.

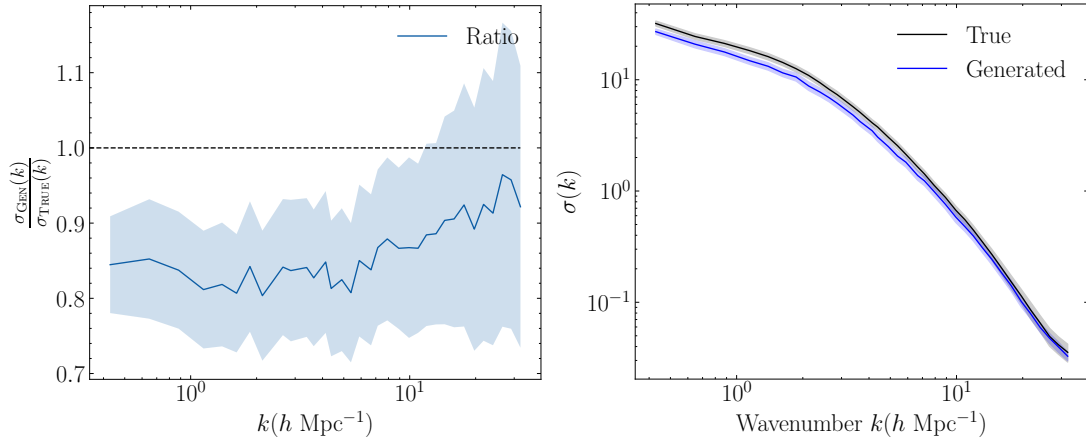


Figure 3.4: Generated ‘CV’ fields. **Left:** Ratio of the standard deviations of the generated fields to that of the true fields. **Right:** Standard deviations of the power spectra in each k bin for the generated and true fields.

have a slight excess correlation around $5 \text{ } h\text{Mpc}^{-1}$.

The standard deviations of the distribution are also consistent although the standard deviations of the generated power spectra are slightly underpredicted relative to the true power spectra at the largest length scales ($k < 5 \text{ } h\text{Mpc}^{-1}$). To construct the standard deviation estimator, we use the jackknife-approach to generate a distribution of 405 estimates of the standard deviation for each set, where the i^{th} estimate corresponds to the standard deviation computed using all samples excluding that of the i^{th} field. We then compute the mean and the standard deviation for these sets in order to capture the mean and the standard deviation on the estimate of the standard deviation in the upper right panel of Figure 3.3. We use the ratios of the jackknife-estimated means and errors in the lower right panel.

The ability to capture the full diversity of k modes for a single parameter may be in tension with the ability to distinguish between and appropriately modulate the power spectrum for different parameters. This tension is enhanced for our assessment of mode collapse in spectral space, compared to canonical machine learning datasets involving discrete classes, since the cosmological parameters that we condition the diffusion generative model on lie on a continuum. Thus, for a given conditioning parameter, a generated field with too much or too little power on certain scales relative to the mean of the statistic is more likely to wander into the typical set of a field with a different cosmological parameter since modulating the parameters also modulates the power spectrum (as in Figure 3.2). In the context of natural images, one often deals with categorical descriptors, where the boundaries between whether an object qualifies as one class or another are typically more clearly delineated, and the ability to capture the full diversity of cats is unlikely to cause the model to ‘wander’

into islands of image space corresponding to a dog or an airplane. Thus the slightly lower standard deviation can be partly attributed to the possibility that the model chooses to compromise on the diversity of samples for a single parameter, in order to be able to accurately generate fields that look different for different parameters.

We compute the covariance between the modes of the power spectrum. With the full covariance we can now statistically quantify the consistency of the generated samples relative to that of the true samples using the multidimensional chi-squared statistic. We use 350 samples of the true fields to set up the reference distribution used to compute the covariance. We compute the inverse covariance and adjust for the Hartlap factor (Equation C.2)⁹⁰. We then compute the multidimensional chi-squared statistic of the entire sample of the true and generated fields, and compute the means of the 405 chi-squared statistics for each distribution following Equation C.3. For the true fields, the mean of the chi-squared distribution is 33.4 while that of the generated is 36.6. Since our power spectra consist of 35 log-spaced bins, this is consistent with the expected mean for a chi-squared distribution with 35 degrees of freedom, i.e. 35.

3.4 PARAMETER INFERENCE

A trained diffusion model can be used to estimate the variational lower bound (VLB) of the log likelihood^{114,56}. In the case of a conditional diffusion model, this likelihood estimate is also conditional, i.e.

$$\begin{aligned} \mathbb{E}_q[-\log p_\varphi(x_0|\theta)] &\leq \mathbb{E}_q[-\log p_\varphi(x_0|x_1, \theta) + \\ &\sum_{t \geq 1} \text{D}_{\text{KL}}[q(x_t|x_{t+1}, x_0) || p_\varphi(x_t|x_{t+1}, \theta)] + \\ &\text{D}_{\text{KL}}[q(x_T|x_0) || p(x_T)]] \simeq L_{\text{VLB}} \end{aligned} \quad (3.2)$$

$$L_{\text{VLB}} = L_0 + L_1 \dots L_{T-1} + L_T \quad (3.3)$$

$$\hat{L}_{\text{VLB}} = \sum_{t < T_{\text{MAX}}} L_t \quad (3.4)$$

where φ denotes the diffusion model architecture, θ is the conditioning cosmology (in our case, a vector with Ω_m and σ_8), p_φ are the reverse (learned) distributions, and q are the forward (analytical) distributions. During training, the diffusion model’s noise prediction loss terms are equivalent to terms of the reweighted VLB¹¹⁴. Since the predicted noise is conditional on cosmology, the terms of the variational lower bound thus encode dependencies on the cosmological parameters. The contrast between the VLB evaluated at one parameter θ_1 relative to another parameter θ_2 for a *fixed* field x_0 can thus be used to find the region of parameter space that maximizes the conditional likelihood for the field.

3.4.1 EXAMINING THE INFLUENCE OF DIFFERENT TIMESTEPS

Since using all the terms L_t that contribute to the VLB is computationally expensive, we first investigate how sensitive each of the contributing terms L_t is to changes in cosmological parameters over a

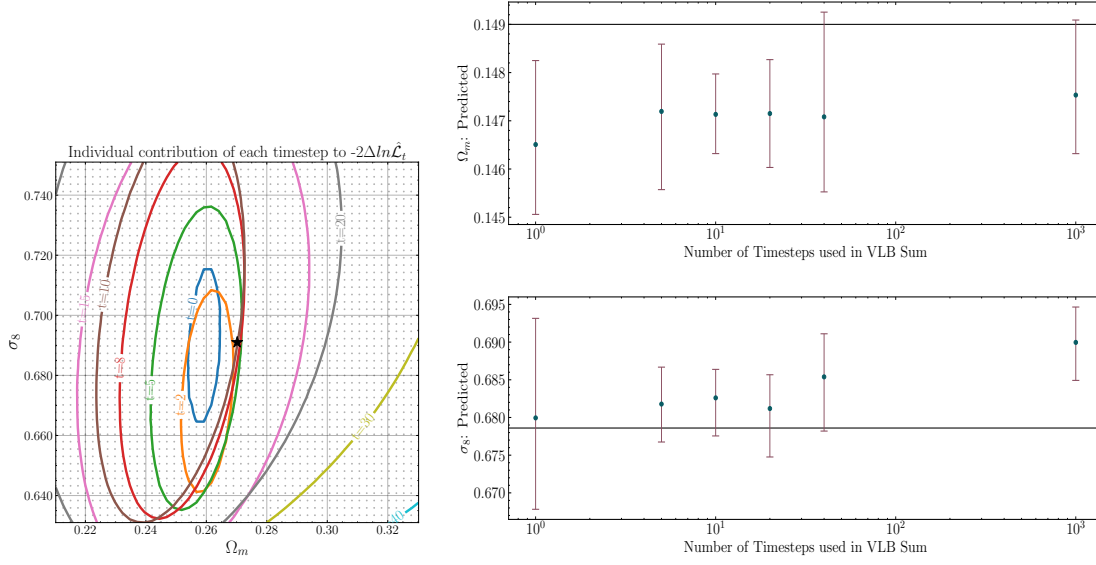


Figure 3.5: Investigating the contribution of the timesteps used in the VLB sum. **Left:** One sigma contours of each $-2\Delta\ln L_t$'s individual contribution for different timesteps. The contours are all centered near the true parameter (black star) and become wider as t increases. **Right:** Mean, and 1 sigma predictions for a single field as a function of the number of timesteps used in the VLB sum optimized by the HMC. Reducing the number of timesteps used to compute the VLB does not significantly affect the diffusion model predictions.

grid in the left column of Figure 3.5. For an input field x_0 and a single seed, we evaluate each of the $L_t(x_0|\vec{\theta}_{\text{EVAL}})$ terms over a 50×50 grid in $[\Omega_m, \sigma_8]$, centered on the value of the true field $\vec{\theta}_{\text{TRUE}}$ and extending to ± 0.06 about the true parameter. To disentangle the *individual* contribution of each term, we subtract the minimum value of each $L_t(x_0|\vec{\theta}_{\text{EVAL}})$ on the grid, and multiply it by 2 to yield $-2\Delta\ln\hat{\mathcal{L}}_t$. We plot the contour corresponding to 2.30, or the one-sigma contour for a chi-squared distribution with two degrees of freedom. Two observations are apparent: all contours are minimized in the vicinity of the true parameter, and the curvature of the contours decreases with increasing time. Increasing time involves increasing the amount of noise added to the input image, which can explain the increased uncertainty in the true value of the parameter. Thus dropping the terms

corresponding to the higher timesteps in Equation 3.2 is unlikely to result in weaker constraints on cosmology.

3.4.2 HMC-BASED PARAMETER INFERENCE

To compute the parameter estimates for fields, we draw samples from the posterior on the parameter using the Hamiltonian Monte Carlo (HMC) method. The Hamiltonian Monte Carlo, or Hybrid Monte Carlo method^{69,148,25} is a Markov Chain Monte Carlo (MCMC) approach that draws samples from a probability distribution $\pi(\theta)$ via the introduction of an auxiliary momentum variable $\mathbf{p}$ and solves the equations of Hamiltonian dynamics in order to update the momentum and position (θ). HMC enables a more efficient exploration of high dimensional probability distributions. In our case, using an HMC also helps us circumvent the problem of having to redefine the extents and granularity of the parameter grid depending on how confident the constraints for a given T_{MAX} are. The Hamiltonian governing the dynamics in the HMC chain is:

$$H = U + K \text{ where } U = -\log \pi(\theta) \text{ and } K = \frac{\mathbf{p}^{-1} \mathbf{M}^{-1} \mathbf{p}}{2} \quad (3.5)$$

$$\begin{aligned} \log \pi(\theta) &= \log p_{\phi}(x_0|\theta) + \log p_{\text{PRIOR}}(\theta) \\ &\simeq -\hat{L}_{\text{VLB}} + \log p_{\text{PRIOR}} = - \sum_{t < T_{\text{MAX}}} L_t + \log p_{\text{PRIOR}} \end{aligned} \quad (3.6)$$

We used the `Hamiltonch`⁴⁸ package for the HMC. We explain more details about the HMC implementation in Appendix C.1. The prior is chosen to be a flat prior over $\Omega_m \in [0.1, 0.5]$ and

$\sigma_8 \in [0.6, 1.0]$. The initial parameter value is always the fiducial value of $[0.3, 0.8]$ and we designate the first 100 samples as burn-in samples that are discarded. We reduce the averages over random noise in each L_t term in Equation 3.2 to a single stochastic estimate of L_t in Equation 3.4.

We now examine the effect of truncating terms in Equation 3.4 using the HMC based parameter inference. Truncating terms allows us to perform inference faster and can allow us to explore the tradeoff between dropping terms for speed and higher precision with more timesteps. In the right column of Figure 3.5, for a single field, we plot the mean predictions and the 15.9 – 84.1 (1 sigma) percentiles computed using 200 samples with the approximate $-\log p_\varphi(x_0|\theta)$ using Equation 3.4 as a function of T_{MAX} on the x axis for the same field. The true value of the cosmological parameter for this field is demarcated by the solid black lines. For this field and parameter, using more terms asymptotically removes the bias on Ω_m while increasing the bias on σ_8 . However, using the first 20 timesteps only changes the mean prediction for Ω_m and σ_8 by -0.26% and -1.27% of the prediction using all 1000 timesteps respectively.

We now turn our attention to the performance of our parameter inference approach relative to a power spectrum baseline. For subsequent HMC-based parameter inference in this section, we use Equation 3.4 with $T_{\text{MAX}} = 20$ to approximate the conditional negative log likelihood. Drawing 500 samples with $T_{\text{MAX}} = 20$ takes ~ 32 minutes for a single field.

To assess the constraining power of the power spectrum, we use the neural posterior estimation approach and train a normalizing flow to represent the posterior of $p_{\text{NF}}(\theta|Pk \cdot \text{Mean})$ using the Lampe¹⁹² package. Since computing the overdensity power spectrum involves dividing by the mean of the field, we further concatenate the mean of the log field as an additional feature, since the pre-

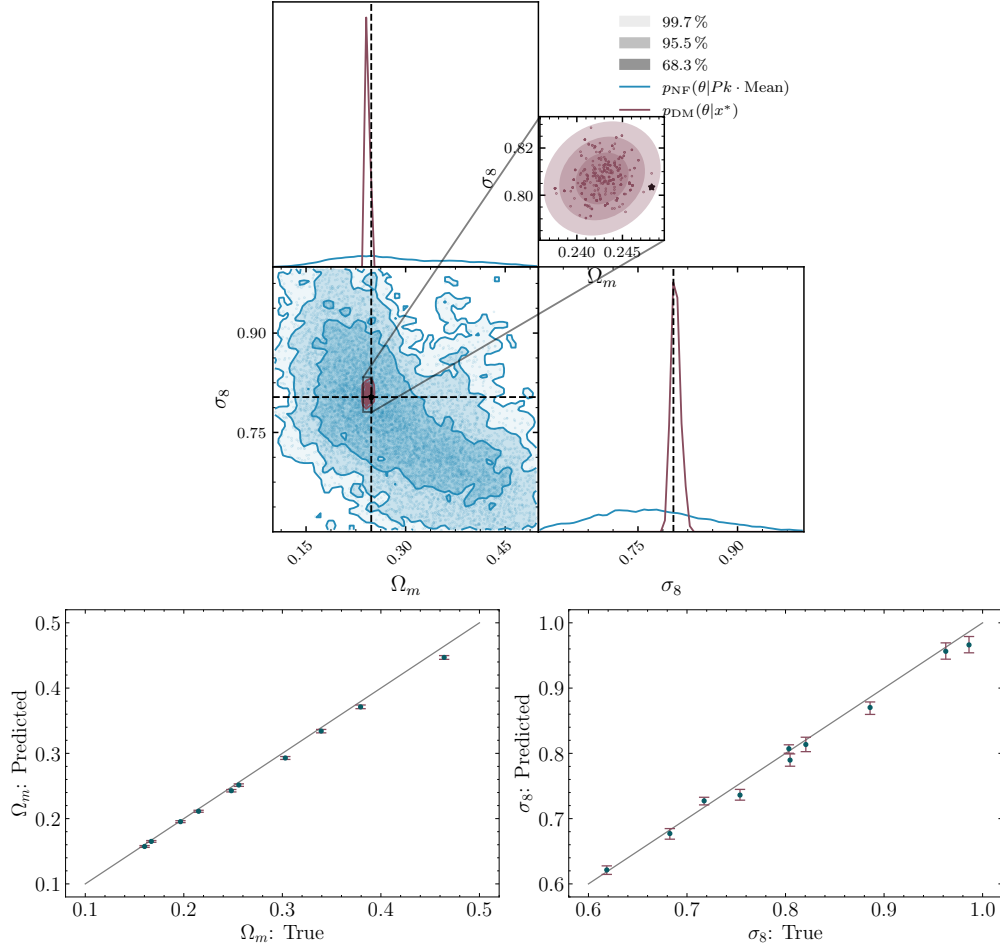


Figure 3.6: Top: Comparison of the parameter estimates for a single field with the power spectrum based estimator and with the diffusion model likelihood. The star demarcates the true parameter corresponding to the input fields. The shaded contours demarcate the 68.27%, 95.45% and 99.73% confidence intervals. **Lower Panel:** Predicted parameter and ground truth parameter for 10 different fields. The error bars correspond to the 15.9 – 84.1 percentiles for the marginal probability distributions for each parameter for each input field.

diction for Ω_m for a small box size of $25 \text{ Mpc}.h^{-1}$ is very sensitive to the mean of the fields. For a single field, we compare the posteriors obtained by drawing 10000 samples from the power spectrum based estimator and 400 samples from the diffusion model+HMC in Figure 3.6. Other details about our implementation are in the Appendix C.1.2.

The diffusion model has significantly narrower constraints for the parameters relative to the power spectrum baseline. Note that, as shown in Figure 3.2, the cosmological parameters Ω_m and σ_8 are strongly correlated at the level of the power spectrum on the small scales probed by our simulations, since they are both modulating the amplitude of the power spectrum. This result is consistent with Ref. ²³¹ who also found that the galaxy power spectrum in conjunction with a multilayer perceptron yielded weak constraints on Ω_m and σ_8 .

In the second row, we plot the predicted cosmological parameters relative to the truth for 10 different fields across different parameters in the validation set. The dots annotate the mean of the 400 samples and the error bars correspond to the 15.9–84.1 percentiles of the samples. The mean of the samples is close to the true value of $\vec{\theta}_{\text{TRUE}}$ over a broad range of parameters. The Ω_m predictions are biased by -0.006 (2% of the true prediction) on average over the range of parameters while the mean absolute bias for σ_8 is 0.01 (1.26% of the true prediction). The σ_8 prediction uncertainties (mean z score $|\bar{z}| = 1.14$) are better calibrated relative to the uncertainties for Ω_m ($|\bar{z}| = 2.75$), but overall, we find the error bars to be slightly under-predicted. We attribute the mis-calibration to the truncation of terms. It is possible that including more timesteps, averaging over more seeds and examining alternate choices of the variance schedule could yield better calibrated uncertainties; we leave this exploration for future work.

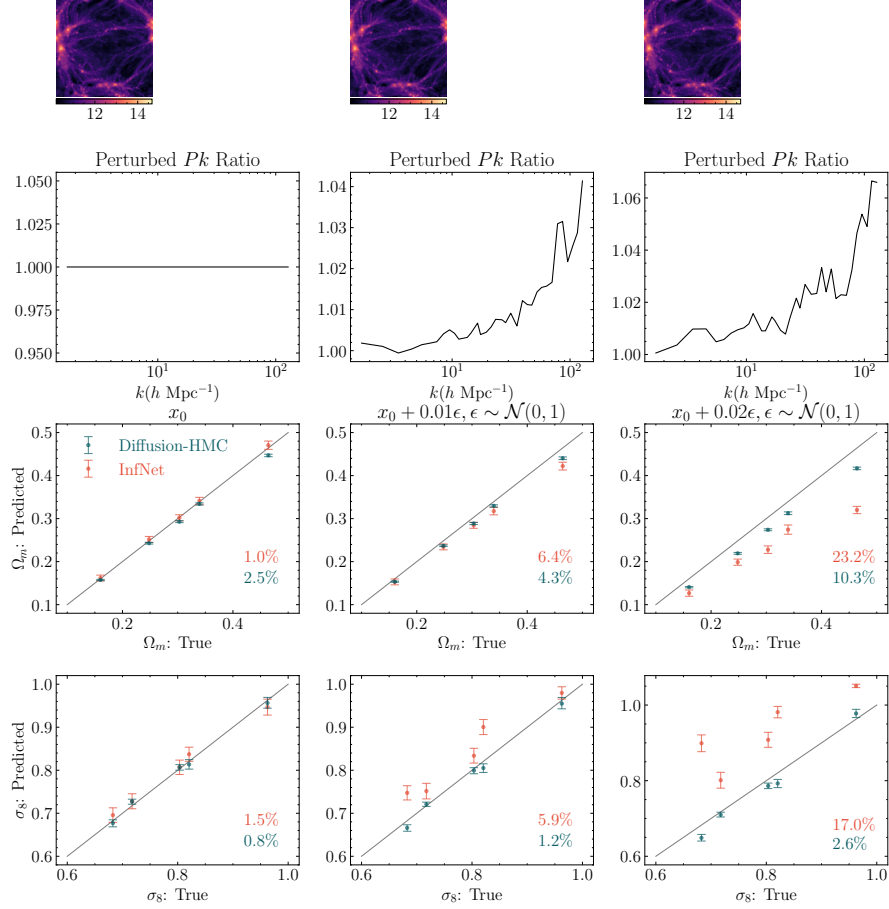


Figure 3.7: First row: An example field corresponding to $\Omega_m, \sigma_8 = 0.33, 0.96$ with its perturbed counterparts, along with the ratio of the perturbed linear field's power spectrum to that of the original field $\frac{Pk_{Noisy}}{Pk_{Original}}$. Second, Third rows: Mean and 15.8 – 84.1 percentile predictions using the Diffusion-HMC estimates (green) and the mean and sigma predictions obtained from the parameter inference network in Ref. ²²⁸, denoted by InfNet. The mean absolute percentage bias (of the true prediction) for each parameter is also indicated. The Diffusion-HMC constraints are more robust to perturbations to the input image.

3.4.3 ROBUSTNESS

Although our constraints are much tighter than those derived from the two point correlation function, robustness to noise and observational systematics is an important consideration guiding the use of parameter inference methods on survey data. Since the terms of the VLB (and the noise prediction loss terms) include terms where the original image has been noised, we hypothesize that the parameter estimates learned by the model are *naturally* more robust to perturbations involving the addition of scaled white noise to the field.

To test this, we examine the extent to which our parameter estimates change relative to the predictions of the parameter inference network in Ref. ²²⁸. The neural network in Ref. ²²⁸ is trained to predict the mean and the standard deviation of the parameters given an input dark matter field. In the leftmost column of Figure 3.7, we compare sample predictions obtained from the diffusion model (Diffusion-HMC) relative to those obtained from the parameter inference network (InfNet) on 5 fields (x_0). In the next two columns, we add increasing levels of $\mathcal{N}(0, \sigma)$ noise with $\sigma \in [0.01, 0.02]$, and examine the estimates for both approaches. We perform this experiment in a noise-agnostic setting, i.e. we assume that we do not know the level of noise a priori and do not modify either the Diffusion-HMC approach or the parameter inference network based approach. We use the same $T_{\text{MAX}} = 20$ and use 200 samples for the Diffusion-HMC estimates for all fields. Although the noise perturbations to the field are visually imperceptible, and correspond to a change of less than $\sim 4 - 6\%$ to the power spectrum at the smallest scales, the predictions of the parameter inference network are significantly ($\sim 6 - 23\%$) disturbed. Indeed, we find that the Diffusion-HMC based

constraints are perturbed far less than the InfNet constraints. The Diffusion-HMC σ_8 constraints are also more robust than its Ω_m constraints.

3.5 CONCLUSION

In this work, we used a diffusion generative model to emulate dark matter density fields conditional on cosmological parameters. The model learns to reproduce the modulation of summary statistics, such as the power spectra, for changes in cosmological parameters. We additionally assess the extent of mode collapse through the lens of cosmic variance by examining the diversity of the power spectra at the fiducial cosmology.

We then directed our attention to the inverse problem of parameter inference and disentangled the contribution of each term L_t in the expression for the Variational Lower Bound to constraints on cosmological parameters. Our findings reveal that the strength of the constraints decrease as t increases. Ref.⁴⁷ and Ref.¹²⁸ explored this question in the context of using the diffusion model’s conditional VLB terms for classification. Ref.⁴⁷ also found a negative exponential weighting of the terms to be optimum while Ref.¹²⁸ found that intermediate timesteps had the highest accuracy when only a single timestep is used. The difference in which terms contain the most information about the conditioning attribute (cosmology / class) is interesting, and could be partly attributed to the difference in their formulation and weighting of the VLB terms and in how changes in the conditioning vector affect an image in different settings. Modulating cosmology modifies global attributes of the fields, such as their intensity and power spectra, as seen in Figure 3.2, while the in-

formation that distinguishes different breeds of pets from each other tends to be relatively localized in the bounding box containing the animal.

If it is indeed always the case that the timesteps nearest to the image manifold contain most cosmological information, one could in future, swap out our discrete time architecture with continuous time diffusion models²¹⁶, where t is a continuous variable with $t \in [0, 1]$ and prioritize steps that lie near the image manifold or $t = 0$.

These insights motivated us to truncate the diffusion model conditional VLB based approximation for $p_\phi(x_0|\theta)$ by subsampling the terms to only use the first T_{MAX} terms (Equation 3.4). This approximation allows us to backpropagate its gradient and plug this estimate for $p_\phi(x_0|\theta)$ into an HMC and sample the posterior on $p_\phi(\theta|x_0)$. The Diffusion-HMC approach yields tight constraints on cosmological parameters, competitive with a bespoke parameter inference network Ref.²²⁸ trained only to infer cosmology given a field.

In our experiments, we use only a single seed in each HMC step and a $T_{\text{MAX}} = 20$ to speed up inference. In our case, the use of an HMC allowed us to circumvent the requirement of choosing a grid with the appropriate resolution required to resolve the constraints, and eliminates the dependence on the number of points in the grid. However, the Diffusion-VLB estimates could also be used in other settings, that do not entail the use of an HMC. While our approach scaled with the number of timesteps we use in the sum, in contrast, Ref.¹²⁸ scaled with the number of classes in the classification dataset.

Lastly, we demonstrate that the diffusion model likelihood confers the Diffusion-HMC constraints with greater robustness against the addition of small amounts of noise to the input image

relative to the behavior of a parameter inference network. This echoes the behavior of diffusion models in other discriminative tasks such as ^{128,40,171}, where diffusion model based classifiers have been shown to possess higher robustness to adversarial examples or perturbations. This is a pertinent finding since Ref. ⁹⁸ showed that the powerful constraints derived from neural network based parameter inference may often come at the cost of their susceptibility to slight perturbations that the canonical two-point correlation based analyses are impervious to. We find that our diffusion-model based parameter inference approach enables more noise-robust field-level inference. It would be interesting to further explore the differences in inductive biases learned by discriminative networks (e.g.: ^{228,205}) relative to generative models repurposed for discriminative tasks.

The simulation volume of the dataset we work with is still much smaller than the scales mapped by astrophysical surveys. Future work could focus on scaling up to more survey-realistic scenarios involving larger simulation volumes and directly observed tracers. Ref. ⁵⁶ also showed that diffusion generative models that work with point clouds can allow one to emulate and perform cosmological parameter inference with point cloud data. Our exploration into robustness could inform applications to real data, and future investigation could focus on conferring and quantifying robustness against other survey-related and observational noise effects. Alternative formulations of the generative process ^{21,236} could also be relevant to this exercise.

In this work, we demonstrated that a diffusion model can be trained not just to emulate fields, but that it’s likelihood estimate can be adapted to work with the Hamiltonian Monte Carlo framework to derive tight constraints on cosmological parameters. This makes a step toward advancing the use of diffusion model based priors for a range of downstream tasks from image generation and

restoration to inference problems.

3.6 CODE

The checkpoint used for the results in this paper is available on Zenodo: doi: [10.5281/zenodo.13993010](https://doi.org/10.5281/zenodo.13993010).

The code for this project is available at [Diffusion-HMC](#)  with a current copy uploaded to the above Zenodo DOI. We acknowledge [The Annotated Diffusion](#), [Improved-Diffusion](#)¹⁵², [VDM](#)¹¹², [DDPM](#) and Ref.²¹⁶ for code snippets and blocks used in the diffusion model architecture and Ref.²²⁸ for the parameter inference baseline code snippets and networks.

Packages: [Hamiltorch](#)⁴⁸, [Lampe](#)¹⁹², [GNU Parallel](#)²¹⁹.

3.7 ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions).

We are grateful to Yueying Ni, Francisco Villaescusa-Navarro, Core Francisco Park, Andrew K. Saydjari, Justina R. Yang, Yilun Du, Ana Sofia Uzsoy, Shuchin Aeron and many others for insightful conversations.

*The following chapter and associated appendix is an adaptation of Mudur, N., Cui, H., Venugopalan, S., Raccuglia, P., Brenner, M.P. and Norgaard, P., 2025. FEABench: Evaluating Language Models on Real World Physics Reasoning Ability. *Workshop on Mathematical Reasoning and AI, NeurIPS 2024* arXiv preprint arXiv:2504.06260 (CC BY 4.0). Changes include omission of some figures and appendix sections and related adjustments of the text. The title, placement of some figures / tables and their referencing text was updated.*

FEABench: Evaluating Language Models on Multiphysics Reasoning Ability

BUILDING PRECISE SIMULATIONS OF THE REAL WORLD and invoking numerical solvers to answer quantitative problems is an essential requirement in engineering and science. We present FEABench, a benchmark to evaluate the ability of large language models (LLMs) and LLM agents to simulate and solve physics, mathematics and engineering problems using finite element analysis (FEA). We introduce a comprehensive evaluation scheme to investigate the ability of LLMs to solve these problems end-to-end by reasoning over natural language problem descriptions and operating COMSOL

Multiphysics[®], an FEA software, to compute the answers. We additionally design a language model agent equipped with the ability to interact with the software through its Application Programming Interface (API), examine its outputs and use tools to improve its solutions over multiple iterations. Our best performing strategy generates executable API calls 88% of the time. LLMs that can successfully interact with and operate FEA software to solve problems such as those in our benchmark would push the frontiers of automation in engineering. Acquiring this capability would augment LLMs’ reasoning skills with the precision of numerical solvers and advance the development of autonomous systems that can tackle complex problems in the real world. The code is available at [FEABench](#).

4.1 INTRODUCTION

Several works have demonstrated the significant potential of large language models (LLMs) in scientific and mathematical domains^{127,241,93,180,223,119,46}. However, existing work has largely focused on analytical mathematical and scientific reasoning skills or the ability to generate code in general purpose programming languages^{222,106}. Addressing the degree of complexity required in numerical simulation-intensive science and engineering workflows – which requires the composition of scientific reasoning with the ability to operate simulation software – remains an outstanding challenge. Many quantitative tasks that form the cornerstone of these workflows require numerical analysis performed with sophisticated computational modeling software. For example, the development of a smartphone requires detailed modeling of the mechanical, thermal, and electrical behaviors of its

many subcomponents. Finite element analysis (FEA) (eg: Ref. ⁵⁵) software develops approximate solutions to the underlying partial differential equations for a physical system, by building discretizations (or meshes) over geometries. The resulting equations are then solved numerically. The vast real-world relevance of FEA to domains like mechanical, biomedical and aerospace engineering, consumer electronics, manufacturing, and scientific research has given rise to software such as COMSOL Multiphysics[®] ^{52,145}, that are indispensable to modeling complex systems with the interplay of non-trivial geometries, and multiple physical phenomena.

Despite the potential impact, the application of LLMs to engineering simulation tasks like FEA remains largely unexplored. In this paper, we begin to bridge this gap by measuring the ability of LLMs and LLM-agents to build models and solve engineering problems using finite element analysis.

Our contributions are the following:

- We introduce a benchmark intended for LLM and agentic research on engineering simulation, a novel domain for LLM benchmarks. The benchmark consists of (1) FEABench Gold: 15 manually verified problems, in addition to (2) FEABench Large: a larger set of 200 algorithmically parsed tasks. The problems in FEABench Gold are (a) quantitatively *verifiable*, that is, if solved completely and correctly, a desired target value will be computed and exported to a table, (b) manually confirmed to have input problem descriptions that are *self sufficient* and do not omit information necessary to solve the problem (c) manually verified to be *solvable*, i.e. we confirmed that if the steps to model the problem are followed faithfully

in COMSOL Multiphysics® the desired target value is computed. The target values are expected to be largely independent of the modeling software. The LLM’s objective is to read the problem specification and operate COMSOL Multiphysics® by generating a sequence of Java calls to its API that would build the model and compute the target. The skills this requires include (1) code generation in a *low-resource* setting, (2) inferring spatial dimensions and representing objects as compositions of geometrical primitives, (3) making correct and consistent physics reasoning decisions (e.g.: boundary conditions and properties).

- We further define two versions of the tasks in FEABench Gold– **ModelSpecs** and **Plan** , to probe different versions of task complexity.
- We introduce a holistic automated evaluation strategy with intermediate metrics that seek to measure different facets of ‘distance to a correct solution’. We benchmark different SOTA LLMs on their baseline (single-turn) performance with these metrics.
- Finally, we design an interface in which an LLM can interact with the COMSOL Multiphysics® API and with specialized auxiliary functions and build an agent that uses execution feedback to improve its solution over multiple turns.

We selected COMSOL Multiphysics® as the framework for our benchmark because it is extensively used for commercial engineering analysis as well as scientific research and supports a wide range of physics models. The FEA workflow is relatively canonical, the reasoning approach for modeling is similar to other FEA software, and problems typically involve a shared conceptual breakdown into a sequence of steps that involve defining 1) Geometry, 2) Material properties, 3)

Physics, 4) Mesh 5) Numerical Analysis and Solver settings, and 6) Postprocessing (details in Appendix D.4.2).

4.2 DATASETS AND TASKS

FEABENCH GOLD The benchmark problems are derived from tutorials in the Ref.⁵³ Application Gallery and are often based on established validation problems or other sources (eg: Ref.^{137,147}).

The input is a natural language problem description with a specific target quantity that needs to be computed. The problems span a range of real world / mathematical systems including heat transfer in objects, modeling a stock option using the Black-Scholes Equation and eigenfrequency analysis of a quantum dot and a beam. Each entry consists of the following main fields:

- **Model Specifications:** A complete description of the task, including geometry, material properties, physics specifications, initial/boundary conditions, and the output to be computed. This field is intended to be general enough to be relevant to other softwares or approaches, but is unambiguous about details such as material properties.
- **Selection Information:** An engineer would typically identify spatial information like geometric selections (points, boundaries, and domains) using the Graphical User Interface (GUI). We provide this field as a substitute for images for LLMs and agents without the ability to receive visual input from the GUI. This information is valid as long as the agent chooses to construct the geometry in COMSOL Multiphysics® in a manner that is reasonably similar to the construction of the ground truth (GT) geometry.

- **Plan:** Step-by-step instructions to solve the problem using COMSOL Multiphysics®.
- **Target Description:** A brief phrase describing the quantity that needs to be computed. This would typically be the value of a specific quantity such as an eigenvalue, pressure or some other variable at a designated location and / or time.
- **Target Value:** The correct value of the target physical quantity.
- **Ground Truth Code:** Lines of COMSOL Multiphysics® API calls that can be executed to build a model that successfully computes the target value.
- **Model Tree:** Executing COMSOL Multiphysics® calls can be regarded as modifying a tree with certain predefined *branches* such as *geometry* and *physics*. The model generated by executing code can thus be represented in a condensed form as a model tree. This is a high-level lossy representation of a solution path, as the code cannot be exactly recovered from the model tree.

Converting tutorials to verifiable benchmark problems requires ensuring that an artifact can be computed from it, generating inputs and the GT solution and verifying that it computes the correct target value (Appendix D.2).

FEABENCH LARGE We further evaluate SOTA LLMs on a larger dataset consisting of 200 Ref.⁵³ Application Gallery tutorial problems. Since these are algorithmically parsed from tutorials, and most tutorials are for demonstrative purposes, the tasks are not structured so as to export a verifiable numerical artifact. They may instead instruct the user to generate specific plots or compute

tables. The input consists of a field termed ‘**Plan**’, which corresponds to the Modeling Instructions in the tutorial. This specifies explicit instructions (similar in nature to the Plan field in FEABench Gold). We additionally save the ground truth API calls in ‘**Code**’ after running some preprocessing steps on the ground truth API calls, in order to resemble the format of the code in FEABench Gold.

ANNOTATED LIBRARY We additionally generate a set of 768 annotated code snippets, by querying an LLM (Gemini-1.5-Flash) to translate code blocks to natural language summaries. Unlike the previous two datasets described, we do *not* use this for evaluation. This is used to retrieve snippets in our agent system.

4.2.1 TASKS

We propose two task variants for LLMs to solve under FEABench Gold: (1) the **ModelSpecs** task, in which the input description for each problem consists of the Model Specifications and Selection Information fields and, (2) The **Plan** task, in which the input description consists of the Plan field. In both cases the LLM agent is expected to return a solution that should consist of the API calls that solve the problem, similar to **Ground Truth Code**. When executed in the COMSOL Multiphysics® API during evaluation, a correct solution will export a table containing a computed value that should match the **Target Value**. We note that both task inputs i.e. the Model Specifications along with the Selection Information, or Plan are independently self-sufficient problem formulations for the LLM to solve using the API. The **Model Specifications** field most closely resembles a naturally occurring real-world problem description. We conduct experiments primarily

on FEABench Gold, unless specified otherwise.

4.3 APPROACH

We first consider a baseline (non-interactive) setting where the LLM generates code without the ability to execute it. The COMSOL Multiphysics® API provides several ‘physics interfaces’ that describe different physics systems and phenomena. A user or LLM needs to select the correct interface and create and modify features under it, in order to implement the physics of the problem. We consider two prompting strategies for both task variants. (1) The first prompting strategy involves a **One-Shot** prompt where the model is given one full input and solution example. (2) The second strategy, **PhyDoc In-Context**, includes the list of valid physics interfaces and features under them in addition to the one-shot example. To further address the challenge of correctly operating the simulation software, we build an agentic system that interacts with the software, and uses tools to improve its solution.

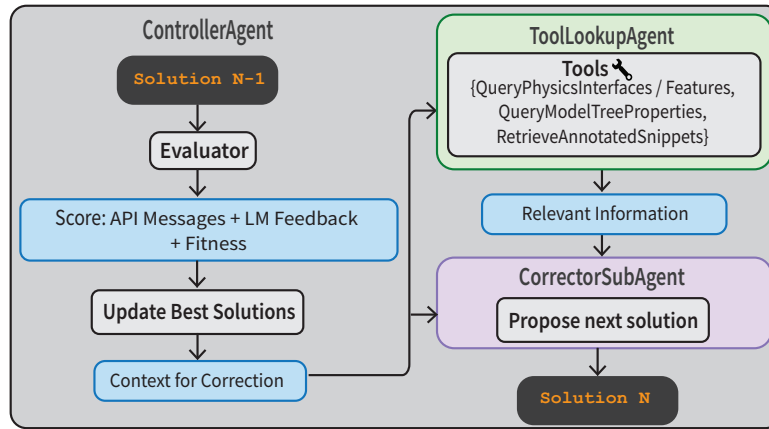


Figure 4.1: An overview of the agent and environment design.

4.3.1 A MULTIPHYSICS REASONING AGENT

Recent work has sought to explore the space of designing optimal Agent-Computer Interfaces^{240,234} primarily for software engineering. However, these frameworks (see Appendix D.1 for a discussion of related work) are tailored to codebase navigation and bash execution: utilities crucial to software development, but of limited relevance to engineering simulation environments. We design a multi-agent system that interacts with the COMSOL Multiphysics® API, as well as tools (or specialized functions). The system is equipped with an `Evaluator` whose feedback is used to compute a ‘fitness’ for each solution. This is used to track the best solutions. The `ControllerAgent` calls a `CorrectorSubAgent` that proposes the next solution given the ‘current’ code and feedback, execution history and the result of tool calls. It in turn delegates tool calls to the `ToolLookupAgent`. To minimize failures or longer-than-desired chains of calls, we adopt an algorithmic sequence of agent calls *except* within the `ToolLookupAgent`. An initial set of 20 samples are generated using `PhyDoc In-Context` and the best solutions are corrected for 20 steps (see Appendix C.2 for details). The best of all solutions is then identified and evaluated. The tool registry consists of a retriever tool `RetrieveAnnotatedSnippets`, a tool that queries the API to return a dictionary of properties under a specified node `QueryModelTreeProperties` and two tools that return the list of valid physics interfaces and features (`QueryPhysicsInterfaces` and `QueryPhysicsFeatures`). We describe our implementation and selection strategy in Appendix D.5, but summarize key features below:

- **LLM-Assisted Semantic Code Search:** To address the challenge of low-resource code gener-

ation, we create a RetrieveAnnotatedSnippets tool that allows the LLM to search the Annotated Library (Section 4.2) for syntactically correct code snippets relevant to a given step (eg: ‘Define the thermal...’ under ‘material’ in Figure 4.2, right panel, ③).

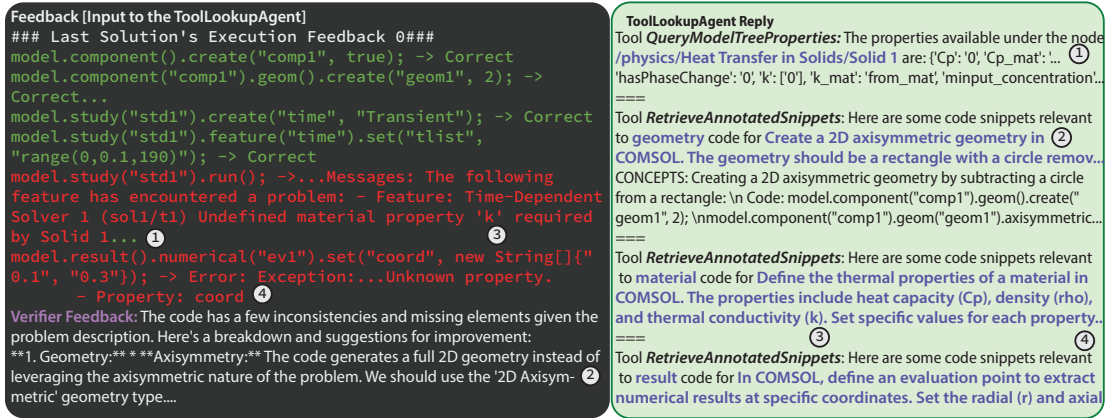


Figure 4.2: The Evaluator’s feedback (left) is passed to the ToolLookupAgent, that calls *tools* and returns their concatenated output. *Violet*, on the left indicates that the Verifier Feedback is returned by an LLM, and on the right, denotes the arguments chosen by the ToolLookupAgent to call the tools with. The numerical annotations highlight the correspondence between the errors and arguments.

- Hybrid Evaluator Feedback: The API returns a ‘reply’ to each line of code in the parsed LLM solution. These messages indicate whether a line of code was correctly executed. We define the percentage of correctly executed parsed lines of code as ‘executability’. API messages alone, however, will not contain information about inconsistencies with the problem description, such as incorrect physical units. Thus, high executability does not guarantee alignment. To address this gap, we call a VerifierLLM to provide feedback (Figure 4.2, left panel) when executability crosses 90%. The API feedback provides a signal on *syntactical correctness* and the VerifierLLM provides a signal on *alignment* and *completeness*.

- **Analytical-Numerical Consistency:** Several problems may allow a scientist to formulate an approximate analytical guess for the target value, even if a precise value may only be derivable numerically. Using this principle, the VerifierLLM sets an analytical guess at the start of the Multi-Turn experiment, given the problem description and compares the numerically computed target with the analytical guess.

4.4 EVALUATION METRICS

Reasoning correctly about the problem and issuing the right calls to operate the API poses a challenging task for even SOTA LLMs. Moreover, a model can only compute a correct target value if it could generate all the code to solve the problem successfully. This renders conventional execution-based code evaluation metrics such as the ‘*pass@k*’ metric^{41,117} challenging to apply to this setting, since most solutions are unable to completely solve the problem. Text-based similarity metrics, on the other hand, are confounded by the preponderance of boilerplate code and the functional equivalence of different code blocks (Appendix D.3.1). To address this challenge, we introduce a multifaceted evaluation strategy that measures correctness even when a target value could not be computed (Table 4.1). Metrics denoted by [†] require execution of the API calls. We delineate the metrics here:

- **Executability[†]:** Executable lines as a fraction of parsed API calls in an LLM solution. A given line may be invalid if it is syntactically incorrect or if it refers to an invalid action (like modifying a property under a non-existent node).

Table 4.1: Summary of Evaluation Metrics

METRIC	ARTIFACTS	SKILLS MEASURED		
		Correctness	Alignment	Physics Reasoning
Executability	API Messages	✓		
Model Tree Score	Model Tree		✓	
Physics Metrics	Physics Code			
Interface Factuality		✓		
Recall Metrics			✓	✓
Feature Dimension		✓		✓
Target Value Metrics	Output	✓	✓	✓

- **Model Tree Score[†]:** Similarity score between the LLM solution’s model tree and a GT tree.

This is normalized so that a solution with no parsed lines of code is scored 0. If it was equivalent to the GT tree, the score would be 1. This measures the *alignment* of the model’s solution path with a successful path.

- **Physics Metrics:** The metrics above analyzed the *entire* solution or its derived artifacts.

The code is a basis to represent the actions the LLM takes to model the problem. Since the physics block is both the most diverse and the most challenging (Figure 4.3), we further evaluated specifically the LLM’s physics actions. The most basic physics action sequence involves: Create Interface (e.g.: HeatTransfer) → Create Feature under Interface (e.g.: TemperatureBoundary) → Modify Feature Properties (e.g.: T0, to set a temperature).

While these metrics offer a granular look into the LLM’s physics reasoning path, some nested physics metrics, such as ‘Feature Dimension’ will not be valid for a problem when there is no overlap between the GT and the LLM code: we mask out these problems while computing

the means for that metric. Our Physics Metrics include

- *Interface Factuality*: What fraction of interfaces created by the LLM are real COMSOL Multiphysics® interfaces and *not* hallucinated?
 - *Interface / Feature / Feature Property Recall*: How many interfaces / features / feature properties created / modified by the GT solution were also in the LLM solution?
 - *Feature Dimension*: For features created by both, does the feature’s spatial dimension match? For example, if an LLM sets a temperature boundary condition on a 1D geometry, this metric checks whether it deduced that the boundary should be 0 dimensional, by comparing the dimension with the GT boundary dimension.
- **Target Relative Error[†]**: We entask an LLM to check that the computed value in the exported table matches the target description and is not a default value, and to parse the response, if so. **Valid Target** is the number of problems for which the LLM judges the value to be valid. We then compute the relative error between the value and the Target (GT) Value. **Relative Error | Strict** computes the mean relative error only over problems for which Valid Target is True, AND the relative error is less than 10%. *Relative Error | Strict is the principal metric one would ideally use to assess whether a problem was truly solved.*

The tables report the means and standard errors on the mean across problems that the experiments were run on. Some nested physics metrics, such as ‘Feature Dimension’ might not be valid for a specific problem, if there was no matching feature between the GT and the LLM code: we mask out these problems while computing the means for that specific metric.

4.5 RESULTS AND DISCUSSION

Comparison across LLMs at baseline. Three closed-source LLMs – Claude-3.5-Sonnet¹³, GPT-4o¹⁵⁴ and Gemini-1.5-Pro¹⁷⁹ – and three open-weights LLMs from the Gemma family^{49,76} are tested on the **ModelSpecs** task under FEABench Gold, with a one-shot prompt (Table 4.2 and 4.3). We find that closed-weight models are able to generate code with moderate executability $\sim 0.60 - 0.79$, implying that LLMs are familiar with the higher-level grammar and syntax of COMSOL Multiphysics® API calls or can infer it from the one-shot example. Getting more granular choices correct is more challenging: LLMs are prone to hallucinating the interface choice (factualty between $[0.54-0.85]$). The open-weights LLMs generally perform worse, especially on the alignment-probing metrics like the Model Tree Score and Physics Recall*. We also compare the performance of the closed-source LLMs on the 200 problems in FEABench Large (Table 4.4). Unlike the human-verified FEABench Gold, FEABench Large instances do not have a single final target artifact, so we only evaluate these against metrics that don’t require execution. *Claude 3.5-Sonnet consistently has the best performance across metrics on both benchmarks.*

4.5.1 ANALYZING FACTORS THAT CONTRIBUTE TO COMPLEXITY

To analyze sources of difficulty in a single-query, non-interactive setting, first, we use the two task versions to decouple whether the bottleneck lies in making correct reasoning decisions or in translating explicit natural language steps into syntactically correct code. Next, we examine gains from

*Since these LLMs were unable to define matching features (feature recall), the dimension metric could be evaluated for fewer than 5 problems and was thus omitted in Table 4.3.

Table 4.2: Code Metrics: Comparison on **ModelSpecs** across LLMs.

Experiment	Executability	Model Tree Score	Valid Target
Claude 3.5 Sonnet	0.79 ± 0.03	0.69 ± 0.07	1/15
GPT-4o	0.78 ± 0.03	0.56 ± 0.06	0/15
Gemini-1.5-Pro	0.60 ± 0.05	0.46 ± 0.07	0/15
Gemma-2-27B-IT	0.56 ± 0.05	0.47 ± 0.07	0/15
Gemma-2-9B-IT	0.44 ± 0.06	0.28 ± 0.06	0/15
CodeGemma-7B-IT	0.52 ± 0.07	0.35 ± 0.06	0/15

Table 4.3: Physics Metrics: Comparison on **ModelSpecs** across LLMs.

Experiment	Interface Factuality	Interface Recall	Feature Recall	Feature Property Recall	Feature Dimension
Claude 3.5 Sonnet	0.85 ± 0.10	0.71 ± 0.13	0.80 ± 0.10	0.22 ± 0.10	0.95 ± 0.05
GPT-4o	0.79 ± 0.11	0.64 ± 0.13	0.55 ± 0.12	0.22 ± 0.11	0.95 ± 0.05
Gemini-1.5-Pro	0.54 ± 0.14	0.43 ± 0.14	0.39 ± 0.10	0.15 ± 0.09	0.86 ± 0.14
Gemma-2-27B-IT	0.69 ± 0.13	0.50 ± 0.14	0.14 ± 0.08	0.11 ± 0.07	-
Gemma-2-9B-IT	0.70 ± 0.15	0.43 ± 0.14	0.06 ± 0.04	0.07 ± 0.07	-
CodeGemma-7B-IT	0.45 ± 0.13	0.21 ± 0.11	0.17 ± 0.09	0.07 ± 0.07	-

Table 4.4: Comparison across models on FEABench Large.

Experiment	Interface Factuality	Interface Recall	Feature Recall	Feature Property Recall	Feature Dim.
Claude 3.5 Sonnet	0.68 ± 0.03	0.50 ± 0.03	0.49 ± 0.03	0.29 ± 0.02	0.96 ± 0.01
GPT-4o	0.66 ± 0.03	0.48 ± 0.03	0.26 ± 0.03	0.20 ± 0.02	0.82 ± 0.05
Gemini-1.5-Pro	0.57 ± 0.04	0.28 ± 0.03	0.44 ± 0.03	0.20 ± 0.02	0.72 ± 0.04

changing the prompting strategy. Finally, we leverage the common structure of the code across problems to examine which block is the most challenging. We fix the LLM to Gemini-1.5-Pro in

subsequent COMSOL Multiphysics®-based experiments. We include a detailed qualitative analysis of a single solution in Appendix D.6.

LLMS FIND IT CHALLENGING TO TRANSLATE PHYSICS REASONING DECISIONS TO CODE. We examine whether the **Plan** task is easier than the **ModelSpecs** task. The comparison between task versions is of interest since both demand slightly different skills. **ModelSpecs** requires the *composition* of planning and reasoning about engineering decisions with translation to valid API calls. Eg: In one heat transfer problem, the LLM needs to infer that the correct representation of a cylinder’s 2D cross-section is a rectangle. The **Plan** task explicitly describes all steps to be followed in natural language and requires the LLM to only *translate* them to valid calls. The comparison between the two tasks offers a way to decouple the difficulty arising from making correct modelling decisions from translating the decisions into calls with the correct syntax. If an LLM or a user fared poorly at making correct modelling decisions but could reliably translate natural language instructions to API calls, it would find **Plan** an easier task. However, we find that *providing an explicit plan doesn’t consistently boost performance on FEABench Gold*. We hypothesize this could be due to the LLM hallucinating API calls by following natural language instructions verbatim. For instance, for Heat Transfer problems, the natural language instructions in **Plan** instruct the LLM to construct a ‘Heat Transfer in Solids’ interface. However, the syntactically correct interface name is `HeatTransfer`. This is also observable in the slight drop on Interface Factuality between the two tasks in Table 4.6.

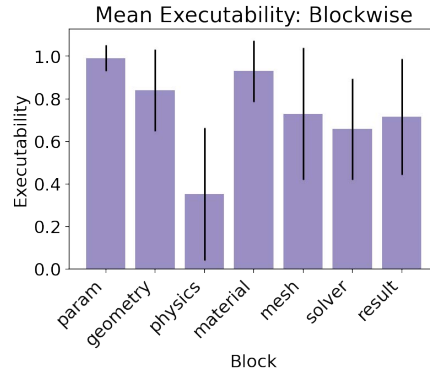


Figure 4.3: Block-wise executability across 300 samples of code and Gemini-1.5-Pro. The physics block has the lowest executability. Error bars denote standard deviations.

GROUNDING THE LLM WITH API INFORMATION BOOSTS PERFORMANCE. The comparison between task formulations indicates that correctly translating decisions to code is a larger bottleneck for our dataset than making correct decisions. We now assess performance, with the list of physics interfaces and features included in the prompt (PhyDoc In-Context). This helps performance, particularly reducing interface hallucinations (factuality: **ModelSpecs** : 0.54→1.0, **Plan** : 0.38→0.85).

PHYSICS SPECIFIC BLOCKS ARE THE MOST CHALLENGING. Figure 4.3 analyzes executability across LLM solutions by breaking down line-wise executability by the block of code the line belongs to. We used the initial set of 20 samples for the 15 problems with the PhyDoc In-Context prompt from the Agent experiment. *The physics block has the lowest executability* with a single query. This motivates our focus on evaluation metrics that focus on the physics block and tools that seek to help ground the LLM’s code with physics-specific information.

Table 4.5: Code Metrics: Comparison across tasks, prompts and agents.

Experiment	Executability	Model Tree Score	Valid Target
ModelSpecs : One-Shot	0.60±0.05	0.46±0.07	0/15
ModelSpecs : PhyDoc In-Context	0.62±0.05	0.58±0.07	1/15
ModelSpecs : Multi-Turn Agent	0.88±0.03	0.56±0.08	2/15
Plan : One-Shot	0.54±0.03	0.39±0.03	0/15
Plan : PhyDoc In-Context	0.59±0.05	0.59±0.06	0/15

Table 4.6: Physics Metrics: Comparison across tasks, prompts and agents.

Experiment	Interface Factuality	Interface Recall	Feature Recall	Feature Property Recall	Feature Dim.
ModelSpecs : One-Shot	0.54±0.14	0.43±0.14	0.39±0.10	0.15±0.09	0.86±0.14
ModelSpecs : PhyDoc In-Context	1.00±0.00	0.71±0.13	0.48±0.10	0.08±0.07	0.59±0.16
ModelSpecs : Multi-Turn Agent	0.93±0.07	0.79±0.11	0.75±0.09	0.24±0.10	0.89±0.07
Plan : One-Shot	0.38±0.14	0.36±0.13	0.43±0.11	0.32±0.11	0.79±0.15
Plan : PhyDoc In-Context	0.85±0.10	0.57±0.14	0.47±0.11	0.13±0.07	0.93±0.07

4.5.2 AGENT RESULTS

Our results underscored the need to ground the LLM’s responses with feedback from and documentation about the API. The interactive Multi-Turn Agent has the best performance of all the COMSOL Multiphysics® experiments on the **ModelSpecs** task on several metrics including executability ($0.62 \rightarrow 0.88$). Although **Relative Error | Strict** is the principal metric one would ideally optimize for, we do not report means over that metric here since the LLM was only able to pass the ‘Strict’ cut (i.e. compute a valid target that was also within 10% of the correct answer) for a single problem in the Multi-Turn Agent and ModelSpecs + PhyDoc In-Context experiments. For this problem, the correct target value is 18.3° Celsius, and the value exported by the LLM is 20° Celsius (specifically $19.999...^{\circ}$ Celsius), which is a default temperature in COMSOL Multiphysics®: this is an indicator of the solution not being solved correctly. While a stricter relative error threshold would eliminate such serendipitous matches, this risks filtering out problems in which a solution might be conceptually correct but differs from the target because of say, differences in numerical choices.

In Appendix D.7, we additionally examine how well the problems can be solved in Python by ‘SWE-agent’, a software-engineering agentic framework²⁴⁰. In this setting, generating executable code is no longer a challenge. However, the bottleneck now lies in achieving the desired precision, due to the absence of verified physics modules and automatic numerical solvers in Python, which requires a user / LLM to explicitly define all equations from scratch. The agent is able to compute a valid target for 11 problems but is only able to compute a solution that passes the ‘Strict’ cut for 4 problems.

4.6 CONCLUSION AND FUTURE DIRECTIONS

FEABench addresses a key gap in the application of LLMs to scientific disciplines by analyzing their ability to operate engineering simulation software to solve problems that require numerical analysis to model mathematical and physical systems. The combination of capabilities required include low-resource code generation and physics and spatial reasoning skills. The complex compositional nature of this task makes it a novel testbed to measure the ability of agentic approaches to interact with a simulation environment and master a domain-specific language well enough to solve real-world quantitative problems. We designed a multiphysics reasoning agent with specialized tools and hybrid feedback to enhance the ability of LLMs to generate executable code. By introducing a multifaceted evaluation strategy and different task formulations, we analyzed the bottlenecks to succeeding at the tasks. Addressing these challenges would advance the development of agentic systems for engineering modelling and simulation.

A way to further increase the complexity of the benchmark could involve more intricate geometries, imported Computer-Aided Design (CAD) models, and requiring the LLM to operate the software via its graphical user interface. While datasets such as FEABench Large provide a useful statistical signal on the quality of code solutions generated across a large number of problems, adding more human verified problems would be valuable. Using an LLM-annotated corpus to boost code executability might facilitate code generation in other low-resource domain-specific language contexts. Conversely, code generation approaches for other low-resource languages³⁷ might reduce the bottleneck of translating predefined decisions to code.

The ability to operate engineering simulation software to quantitatively analyze a problem would augment LLMs’ reasoning skills with the software’s numerical precision and inbuilt checks, and significantly push the ceiling on the complexity of problems that LLMs can accurately and reliably solve. Unlocking this potential would bring LLMs a step closer to being able to serve as grounded ‘engineering assistants’ that can autonomously run precise simulations to innovate and optimize designs and answer quantitative questions about physical phenomena in the real world.

4.7 ACKNOWLEDGEMENTS

We are grateful to Eser Aygün for valuable suggestions on agent design. We thank Stephan Hoyer and Marc Coram for useful comments on the draft, and are grateful to Rachel Stigler for guidance. At Harvard, NM is partially supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions).

The following chapter and associated appendix is based on ongoing work involving Nayantara Mudur, Carolina Cuesta-Lazaro, Michael Toomey and Douglas P. Finkbeiner and appeared at the Workshop on LLM for Scientific Discovery: Reasoning, Assistance, and Collaboration (LM4Sci), Conference on Language Modeling (COLM), 2025.

<https://openreview.net/forum?id=xnPJOtPmK3> (CC BY 4.0) *This version includes an additional appendix experiment, an expanded introduction and example. The placement of some figures / tables and their referencing text was updated.*

5

An LLM-driven framework for cosmological model-building and exploration

OUR UNDERSTANDING OF THE EVOLUTION of the Universe relies on the existence of dark energy and dark matter—components whose true nature remains a mystery, despite accounting for 95% of the Universe’s energy budget. Recent surveys have additionally revealed discrepancies with respect to the canonically accepted standard concordance (Λ CDM) model, potentially hinting at new

physics. Given the success of large language models (LLMs) at completing research-level tasks, we investigate their capability to autonomously propose, implement, and test alternate cosmological theories. We use our framework to challenge Claude Code in three experimental settings: (1) implementing alternative models from curated descriptions by modifying a physics simulation codebase, (2) performing the same implementation directly from a research paper, and (3) generating novel hypotheses to better explain recent observations. Across two benchmark models with ground-truth implementations, Claude Code successfully implemented a “Thawing Quintessence” dark energy model. However, it failed to generate correct observables for a more complex “Early Dark Energy” model despite successful code compilation. When working directly from papers rather than curated descriptions, numerical accuracy degraded significantly though qualitative behavior remains correct. Most remarkably, Claude Code’s self-proposed dark energy model achieved a better statistical fit to observations than the standard model, though at the cost of additional parameters.

5.1 INTRODUCTION

Cosmology—the study of how the Universe evolved from the Big Bang to today—faces a double crisis. The Lambda Cold Dark Matter (Λ CDM) model, our current standard model of cosmic evolution, has withstood a barrage of tests in predicting observations from the formation of the first atoms to large-scale galaxy clustering. Its many successes include its ability to predict properties of the Cosmic Microwave Background (CMB) such as the Baryon Acoustic Oscillation (BAO) peaks and imprints of the BAO in galaxy clustering, among others. However, the true physical nature of

dark matter and dark energy – two key components that form the underpinnings of this model – remains elusive. Beyond these theoretical issues lies also an empirical problem: even accepting dark matter and dark energy as given, Λ CDM increasingly struggles to fit high-precision data. Discordant measurements of key cosmological parameters from recent surveys have further raised specters of doubt.

Most notable among these is the ‘Hubble Tension’^{66,204}, a nearly 5 sigma discrepancy between the measurement of the expansion rate of the Universe, H_0 , between model-dependent ‘early-Universe’ inferences involving the CMB and ‘late-time’ probes involving Type Ia supernovae and Cepheids¹⁸⁶. Another source of contention arises from measurements of σ_8 : the amplitude of the matter power spectrum at $8h^{-1}\text{Mpc}$. The S_8^* tension, that currently hovers around 2 – 3 sigma, refers to the disagreement between the CMB-inferred value of S_8 and that inferred from low-redshift growth-of-structure probes involving cosmic shear and galaxy clustering². The most recent cosmic surprise emerged in 2024-25 from the Dark Energy Spectroscopic Instrument (DESI)^{130,64}, whose observations favor a dynamically evolving dark energy paradigm over Λ CDM, in which the dark energy equation of state parameter, w , is -1. Exploring alternative models has thus become an urgent priority.

Resolving these cosmological tensions requires exploring a vast landscape of theoretical models that propose alternative forms of dark matter and dark energy, such as modified gravity theories, interacting dark sectors, dynamical dark energy models etc. Proposing a viable model is also a challenging endeavor: any promising alternative must ideally reproduce the successes of the standard

*This tension is often discussed in terms of the disagreement in the inferred value of $S_8 = \sigma_8 \sqrt{\Omega_m/3}$.

concordance model, provide a better fit to data, and alleviate one or more cosmological tensions. Each theoretical framework introduces new parameters, functional forms, and physical assumptions, creating a high-dimensional space where most regions remain unexplored due to computational and human limitations. Large language models, trained on the vast literature of theoretical physics, could potentially sample from this space of hypothesis systematically and generate novel combinations and extensions that human researchers might not have considered.

Testing cosmological theories requires specialized simulation codes, such as the Cosmic Linear Anisotropy Solving System (CLASS) and Code for Anisotropies in the Microwave Background (CAMB), that take theoretical parameters as input and produce synthetic observational data as output. Implementing an alternative cosmological model involves a canonical workflow in which these standard codebases are modified to accommodate the new physics. As a next step, researchers evaluate Bayesian posteriors on the observables and assess whether the proposed model can explain the data better than Λ CDM does.

Model building in cosmology makes for a compelling yet challenging LLM application for several reasons. Firstly, almost all papers implementing an alternate cosmological model rely on 2-4 codebases. Foremost among these is the CLASS codebase. Few subfields in the sciences have such a *unified workflow*. Secondly, these alternate theories need to pass a set of common tests that evaluate their ability to reproduce observations. From a machine-learning perspective, this means that cosmological model discovery can be reformulated as a setting in which all models are tested against and optimized against a set of *well-defined metrics*. Lastly, beyond implementation, this field offers an intriguing test bed to ask whether recent developments in artificial intelligence could help us discover

new physics. In this work, we introduce a multi-stage framework that establishes a systematic test bed in which to measure the ability of an LLM-driven agent to autonomously propose, implement and explore cosmological models. We measure both the potential of the agentic framework to correctly implement a given cosmological model, as well as propose its own theoretical model.

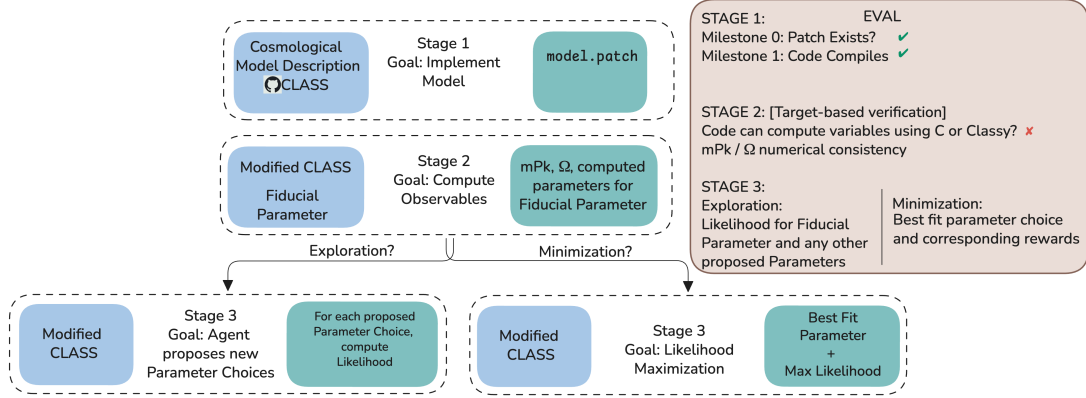


Figure 5.1: Flow diagram depicting the input state for each stage (blue) the output artifacts (green) and the task that the agent has to perform. Once the agent has completed its work, the outputs are automatically evaluated on the basis of both binary milestones that track initial progress as well as numerical metrics. The evaluation box (brown) depicts a template evaluation report for the agent's 'submission'.

5.2 ENVIRONMENT DESIGN

The agent performs its tasks in a containerized code execution environment using a Podman-based emulator for Docker. Containerized code execution enables us to control which files, directories and packages the agent has access to and / or can modify. The environment is designed to eliminate the possibility of the agent accessing the internet to look up information or code (see Appendix E.2 for details).

The agent will modify the physics simulation codebase of CLASS¹²⁶. The CLASS code is avail-

able at https://github.com/lesgourg/class_public. The CLASS machinery is implemented in C, although it can additionally be used to compute observables via a Python wrapper to the C code (Classy). The LLM agent is launched in a directory that contains the base CLASS repository, instantiated as a GitHub repository, in order to easily export the changes as a patch. The repository additionally contains contextual prompts with instructions for each stage, information about CLASS, the cosmological-model specific information as well as auxiliary helper and testing code. We work with the Claude Code¹⁴ agentic framework.

5.2.1 COSMOLOGICAL MODEL SPECIFICATION

Cosmological observables, such as the matter power spectrum, which quantifies the clustering of matter at different scales, are determined by both the underlying theoretical model as well as the specific values of its characteristic parameters. We use $\mathcal{M}$ to denote a theoretical model and $\mathcal{P}$ to denote a specific point in parameter space for the relevant cosmological model.

A model instance typically consists of the following files:

- Model Description [`description.md`]: This is a file containing the description of the cosmological model of interest, $\mathcal{M}$. Appendix E.1.2 includes the description for one of the models we tested the agent on.
- Fiducial Parameter Files: The fiducial parameter set is used to test whether the modified code can actually be used to compute cosmological observables. We refer to this parameter set as

$\mathcal{P}_{Fiducial}$.

- Other problem-specific contextual variables [`problem_context.yaml`]: This is a YAML file that seeks to account for any other problem specific contextual variations and instructions such as the manner in which the agent should explore the models parameter space, as well as the observables that need to be computed and saved.

5.3 STAGES

5.3.1 STAGE 1: MODEL IMPLEMENTATION

In the first stage, the agent’s objective is to implement the cosmological model description by modifying the CLASS code, to enable it to work with the fiducial parameter file. The agent is instructed to read the provided model description. It must then modify code across the CLASS code base and seek to compile the modified CLASS. The core CLASS machinery has over 43000 lines of C code, with another 10000 lines in auxiliary numerical files. To mitigate the challenge of navigating the code base, we added three ‘filemap’ files: high-level summaries of the organization of the ‘background’, ‘input’ and ‘perturbations’ files, in addition to the model description and the fiducial parameter file. The desired state of the environment upon the completion of this stage is code that now successfully implements the alternate cosmological model.

5.3.2 STAGE 2: COMPUTING OBSERVABLES

The agent must now use its modified code to compute observables for the fiducial cosmology, using the provided parameter files. The agent is instructed to use a helper script that computes observables

using the modified code, that should now implement $\mathcal{M}$. These observables are relevant to the target artifact-based evaluation.

If the code fails to execute, another LLM (o1-2024-12-17)¹⁵⁵ is queried to provide feedback comparing the description of the cosmological model with the patch and identify any inconsistencies. This functions as an in-the-loop LLM-as-Judge setup. If the code computes observables successfully, a Pickle file containing these observables is saved, and this serves as the key output artifact relevant to Stage 2. The agent is also instructed to call a module that displays the present day values of key computed cosmological parameters for the $\mathcal{P}_{Fiducial}$ to confirm if the values seem reasonable, as a self-consistency test.

5.3.3 STAGE 3: PARAMETER EXPLORATION

In this stage, the agent is instructed to explore other parameters and compute the log likelihood with respect to observational data for both any new parameters and $\mathcal{P}_{Fiducial}$. In this case, we will compare the theoretical models to observations on how galaxies cluster on large scales at late times. In particular, we use data from the DESI collaboration⁶⁴, and combine its likelihood with the early Universe priors described in Appendix A of Ref.⁶⁴.

We consider two pathways in Stage 3. The ‘*Exploration*’ pathway seeks to mirror the model exploration sections of many theoretical cosmology papers, in which cosmologists manually vary key parameter values to preliminarily assess the model’s behavior in different physical limits and examine its merits. This pathway is also a less prohibitive option if the theoretical model happens to be one that is challenging to optimize, as in the *Early Dark Energy* model. In this pathway, we essentially

ask the agent to propose and explore alternate parameter choices, $\{\mathcal{P}\}$ ‘manually’.

In the ‘*Minimization*’ pathway, the agent must algorithmically find the subset of parameters that best fits the observations. Here, we optimize the parameter file with respect to the likelihood. The best-fit parameter, that maximizes this combined likelihood is subsequently referred to as $\mathcal{P}_{Best-Fit}$.

5.4 EVALUATION

Our setup is designed so that once the agent has completed its tasks its changes are saved in a GitHub patch in a ‘submission’ folder on the host system. We then separately launch another Docker container and run an evaluation script that applies the patch containing the changes the agent made and evaluates its progress on both binary milestones and continuous evaluation metrics. The first 2 binary milestones examine whether it was able to generate a `model.patch` file that produces code that can compile. The next binary milestone checks whether the compiled code can be used to compute observables with $\mathcal{P}_{Fiducial}$. We then numerically compare the matter power spectrum at two different redshifts (a cosmological proxy for time), $z = 0$ (present-day) and $z = 1$, and the evolution of the energy density of the relevant dark energy component as a function of redshift. We report maximum relative deviation with respect to the target repository in Table 5.1. For the evaluation of Stage 3, we compute how well the agent implemented theoretical framework $(\mathcal{M}, \mathcal{P})$ fits data, for all $\mathcal{P}$ explored / derived via optimization.

Table 5.1: Results for Target Artifact-based Evaluation Experiments

Model	Early Dark Energy	Thawing Quintessence	Thawing Quintessence [from paper]
Milestone 0 [Patch Exists?]	✓	✓	✓
Milestone 1 [Code compiles?]	✓	✓	✓
Milestone 2 [Observables can be computed?]	✗	✓	✓
$\Omega_X(z)$ Max Frac. Deviation	-	$4.43e-10$	0.248
$mPk(z=0)(k)$ Max Frac. Deviation	-	$1.31e-4$	0.0997
$mPk(z=1)(k)$ Max Frac. Deviation	-	$1.67e-05$	0.0266
Stage 3 Evaluation	-		
DESI-Planck $\mathcal{P}_{Fiducial} \ln \mathcal{L}$		$-4.21e3$	-47.8
DESI-Planck $\mathcal{P}_{Best-Fit} \ln \mathcal{L}$		-14.7	-14.8

5.5 RESULTS

Unless otherwise specified, we used the `claude-sonnet-4-20250514` model¹⁵ in all experiments.

5.5.1 EXPERIMENTS ON MODELS THAT PERMIT TARGET-BASED EVALUATION

DESCRIPTION OF THE PHYSICAL MODELS

We examined the performance of Claude Code on two cosmological models: in increasing order of difficulty, a phenomenological *Thawing Quintessence*¹⁶⁴ model that mimics the qualitative behavior of (late-time) dark energy, and (2) an *Early Dark Energy* (EDE) model^{110,170,136,108,169}, a particle physics inspired extension to standard cosmology. Both models have ground truth implementations

available (but known to be public only for EDE).

For a successful implementation of *Thawing Quintessence*, the model needs to introduce the correct evolution of the dark energy density as a function of time, denoted as $w(z)$, together with its additional parameters and analytical derivative. In this case, dark energy behaves as a homogeneous fluid and to compute the dark energy evolution the model can use the existing fluid evolution framework in CLASS.

While fluid models with $w(z)$ profiles prescribe how dark energy's pressure-to-density ratio evolves over time, EDE requires solving for a fundamental dynamical variable that must achieve a specific behavior—rapidly rising to 10% of the universe's energy at a precise epoch before quickly decaying—through solving a new set of differential equations rather than simply specifying this evolution. This dynamical variable creates its own fluctuations that interact with matter fluctuations at all scales, requiring coupled equations where the variable's evolution both depends on and influences the universe's expansion rate, unlike fluid models where the energy evolution is given as an input. The implementation complexity arises because we must compute from first principles how this dynamical component evolves and gravitationally affects all other components at every scale, rather than treating EDE as having only aggregate properties that affect the background expansion—making the computational framework substantially more involved than the prescriptive fluid approach used for *Thawing Quintessence*.

The results of the evaluation are in Table 5.1. Whereas the model can successfully implement the *Thawing Quintessence* model, it fails to successfully compute observables for the EDE model. For the EDE model, the agent inconsistently integrates a density parameter for the EDE model with

certain functions despite an implementation that is otherwise promising. This breaks a section of the code that adjusts the density parameter via a numerical shooting algorithm.

The generated code patches can be found in Appendix E.4.

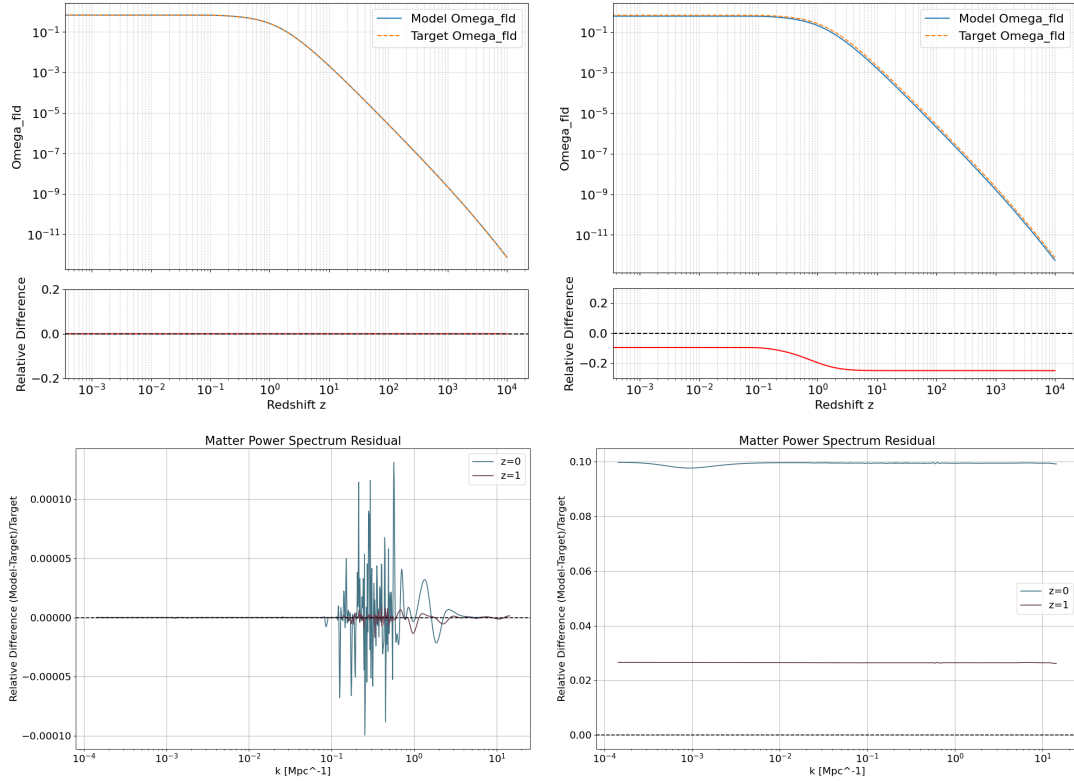


Figure 5.2: Target Artifact-based evaluation: The evaluation for Stage 2 compares the evolution of the relevant dark energy component (Ω_{fld} for *Thawing Quintessence*) and the matter power spectrum at two redshifts against their corresponding ground-truth values generated by a target implementation of the model. While the agent is able to generate executable code when provided with the *paper* (right column), the numerical errors are much higher than when it is given a condensed, yet detailed *description* (left column).

5.5.2 ‘PAPER TO CODE’

For *Thawing Quintessence*, we ran the same experiment after swapping the `description.md` file with the `tex` file of the corresponding paper¹⁶⁴, to examine whether the agent is still able to generate working code given just the paper, as opposed to a well-specified problem description. We still include the fiducial cosmology parameter file in this setting so we can later perform the Stage 2 evaluation against numerical target artifacts. The agent is still able to generate executable code. However, the comparison against the target artifacts, in Figure 5.2 reveals significantly higher numerical inconsistencies relative to the condensed model description. This is in part because the model omits a change to the code where it needs to explicitly rewrite the numerical integration prescription for this model. These differences do not seem to impact the best fit likelihood values of the models (Table 5.1).

5.5.3 HYPOTHESIS GENERATION

Thus far, our experiments measured the ability of the agent to implement a *given* cosmological model $\mathcal{M}$. In this section, we further examine whether it is capable of developing its *own* theoretical model, and implementing it by modifying CLASS. The model is instructed to read an auxiliary prompt and propose its *own* cosmological model, prioritizing the following criteria (in order of importance): (1) novelty, (2) ability to alleviate cosmological tensions and simultaneously providing a good fit with respect to data, and (3) ease of implementation in CLASS. We additionally provide it with an example description derived from the description for *Thawing Quintessence* and an example

parameter file for *Thawing Quintessence*. It must first create its own `description.md` file describing its proposed cosmological model, $\mathcal{M}$, and a corresponding $\mathcal{P}_{Fiducial}$, and then proceed with the remaining stages as before. Since there doesn't exist a target implementation in this setting, the Stage 2 target-based is skipped in the evaluation of the agent's submission.

Table 5.2: Results for LLM-proposed model experiments

Model	Novel One-Shot Run 1	Novel One-Shot Run 2
Milestone 0 [Patch Exists?]	✓	✓
Milestone 1 [Code compiles?]	✓	✓
Milestone 2 [Observables can be computed?]	✗	✓
Stage 3 Reward	-	
DESI-Planck $\mathcal{P}_{Fiducial} \ln \mathcal{L}$		-111
DESI-Planck $\mathcal{P}_{Best-Fit} \ln \mathcal{L}$		-12.6

We ran two experiments in this setting (summarized in Table 5.2). In the first experiment, the agent's proposed model implementation fails to cross the second milestone. In the second experiment, the agent succeeds at generating executable code for its model. The model proposes the phenomenological dark energy model described below (with the full description in Figure 5.3 below):

$$w(z) = w_0 + w_a \cdot \frac{z}{1+z} + A_{\text{osc}} \cdot \exp(-\alpha z) \cdot \sin(\omega \cdot \ln(1+z) + \varphi), \quad (5.1)$$

where z is the cosmological redshift, w_0 is the present-day equation of state, w_a controls the linear evolution with redshift, A_{osc} is the amplitude of oscillations, α is the damping parameter that

causes oscillations to decay at high redshift, ω controls the frequency of oscillations in $\log(1+z)$ space, and φ is the phase offset of the oscillations.

In Stage 3, its $\mathcal{P}_{Best-Fit}$ value yields a likelihood of -12.59. For reference, the $\mathcal{P}_{Best-Fit}$ for Λ CDM, the canonical model, yields a likelihood of -12.88. However, the agent appears to achieve the superior fit by proposing a six parameter extension to Λ CDM (including w_0 and w_a). Adding more parameters to a model should in principle improve fits to data, since the theory now possesses more degrees of freedom.

Figure 5.3: A cosmological model proposed by the LLM.

The Oscillatory Dark Energy (ODE) Model

This is a novel dark energy model that introduces oscillatory behavior in the dark energy equation of state as a function of redshift. The model is designed to address current cosmological tensions, particularly the Hubble and S8 tensions, by allowing for dynamic dark energy behavior that can modify the expansion history and structure formation in ways that static models cannot.

The core innovation of this model is an oscillatory equation of state that experiences damped oscillations around a central value, modulated by the cosmic scale factor. This can potentially provide the right combination of early-time and late-time effects to reconcile observational discrepancies.

The equation of state is parametrized as:

Figure 5.3 (Continued)

$$w(z) = w_0 + w_a \cdot \frac{z}{1+z} + A_{\text{osc}} \cdot \exp(-\alpha z) \cdot \sin(\omega \cdot \ln(1+z) + \varphi)$$

where:

- z is the cosmological redshift
- w_0 is the present-day equation of state (similar to CPL parametrization)
- w_a controls the linear evolution with redshift (similar to CPL parametrization)
- A_{osc} is the amplitude of oscillations (typical range: 0 to 0.5)
- α is the damping parameter that causes oscillations to decay at high redshift (typical range: 0.1 to 2.0)
- ω controls the frequency of oscillations in $\ln(1+z)$ space (typical range: 1 to 10)
- φ is the phase offset of the oscillations (typical range: 0 to 2π)

This parametrization ensures that:

1. At very high redshifts ($z \gg 1/\alpha$), oscillations are damped and $w(z) \approx w_0 + w_a$
2. At intermediate redshifts, oscillatory behavior can modify the expansion history
3. At low redshifts, oscillations persist but with reduced amplitude
4. The model reduces to CPL parametrization when $A_{\text{osc}} = 0$

Figure 5.3 (Continued)

Parameters / .ini Structure for Added Parameters: This model introduces 6 new parameters beyond the standard Λ CDM parameters:

Parameter in .ini	Description	Typical Range (Priors)
w0_fld	Present-day equation of state parameter	$-1.3 \dots -0.7$
wa_fld	Evolution parameter (CLP-like)	$-2.0 \dots 2.0$
A_osc_fld	Amplitude of oscillations in equation of state	$0.0 \dots 0.5$
alpha_osc_fld	Damping parameter for high-redshift oscillations	$0.1 \dots 2.0$
omega_osc_fld	Frequency parameter for oscillations	$1.0 \dots 10.0$
phi_osc_fld	Phase offset of oscillations	$0.0 \dots 2\pi$

Additional background field variables: When running CLASS with this model, the following variables related to the oscillatory dark energy component should be accessible in the `background.dat` output file:

- `w_fld`: The equation of state $w(z) = p_{\text{ode}}/\rho_{\text{ode}}$ of the oscillatory dark energy fluid
- `(.)rho_fld`: Energy density of the oscillatory dark energy component

Implementation Details

- This model will be implemented using the fluid equation of state framework in

Figure 5.3 (Continued)

CLASS

- The oscillatory component will use the existing fluid dark energy infrastructure in `background.c`
- We will modify the `w_fld` computation to include the oscillatory terms
- The model requires careful integration due to the oscillatory nature, so integration tolerances may need adjustment

5.6 CONCLUSIONS

This work introduces an end-to-end framework that seeks to replicate the cosmological model-building research pipeline and examines the ability of an LLM-based agent, Claude Code, to execute multiple stages of the process. We first examine performance on two cosmological models for which ground truth codebase implementations are available: while the agent is able to generate code that compiles for both models, it only crosses all milestones and is able to complete Stage 3 for one. We further performed an ablation test by providing the agent with the text of the paper, instead of the well-specified problem description. This setting raises the challenge of the task since the agent must now also *extract* the most relevant information to modifying the codebase. Lastly, we took a step in the direction of automating theory development by additionally entasking the model to now devise and implement its own cosmological model. While the model succeeds at creating a model that attains a slightly better fit to data, it does so at the cost of introducing additional parameters. Our results demonstrate both the promise and limitations of LLM-driven scientific model discovery,

providing a concrete framework for evaluating AI agents on complex scientific workflows that combine domain knowledge, mathematical implementation, and empirical validation. While our experiments measured agentic ability with minimal human interaction (only to grant permissions), the framework directly accommodates greater human interaction. By accelerating the pace of cosmological model exploration such pipelines may facilitate progress on one of the most pressing questions in modern physics: understanding the nature of dark energy.

Our current framework represents initial progress on the route to automated cosmological model discovery. Moving forward, we plan to expand our approach in two key directions to enable more systematic and principled exploration of the theoretical landscape. First, we will implement evolutionary algorithm-inspired approaches that use the computed likelihood to iteratively guide the discovery of improved dark energy models. Rather than exploring fixed parameterizations, this approach will enable the LLM to evolve functional forms for scalar field potentials and dark energy evolution, using reward signals to steer toward models that better explain observations. We will extend this framework to incorporate interactions between dark energy and dark matter sectors—a theoretically motivated class of models that could address multiple cosmological tensions simultaneously.

Second, we will expand our current likelihood-based reward function to include other rewards, such as the ability to alleviate other cosmological tensions such as the H_0 tension, and compute Bayesian evidence ratios comparing proposed models to Λ CDM. The inclusion of the latter addresses a critical limitation of our current approach: models with additional parameters can achieve better fits simply by overfitting the data. By using evidence ratios as rewards, we will naturally penalize models

for their increased complexity while rewarding simpler extensions that can still explain the observations. This Bayesian framework will provide a more principled basis for model selection and better align our automated discovery process with standard practices in cosmological model comparison. Together, these enhancements will create a more robust system for autonomous cosmological theory development, capable of both broader exploration and more rigorous evaluation of proposed alternatives to our current understanding of the cosmos.

5.7 ACKNOWLEDGEMENTS

We acknowledge helpful discussions with Siddharth Mishra-Sharma, Mikhail Ivanov and Thomas Harvey. This work is supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions).

6

Conclusion

In the years since Chapter 1, other stellar-reddening-based maps used in cosmological surveys such as Ref. ^{44,246} and for detailed studies of the interstellar medium (ISM)⁷⁰ have been released. From the perspective of better observational systematics calibration in cosmology, the availability of multiple maps of Galactic extinction facilitates work such as Ref. ¹⁰⁹, that studied the effect that swapping the choice of Galactic extinction map has on the inferred value of σ_8 and consequently the level of the σ_8 tension. The wealth of new data from the recently commenced Rubin Observatory mission will enable even more detailed stellar-reddening based maps, while potentially exacerbating the problem of star-galaxy separation, because of the ability to now capture fainter, hitherto ‘unseen’ stars. Most three-dimensional dust map reconstructions continue to rely on Gaussian-process based priors. But

the ability of the trained diffusion model in Chapter 2 to generate realistic samples of dust, invokes the fascinating possibility of constructing detailed high-resolution structure akin to those visible in emission-based maps like WISE, using generative priors capable of capturing non-Gaussianity. This may require considerable methods development due to the sheer size and number of data points involved in the dust reconstruction inference problem. Moreover, while most potential sources of training data exists in two dimensions (eg: high-resolution emission maps), three-dimensional maps are of most utility in ISM research.

Moving on to Chapter 3, the next step would involve applying approaches such as those we described here to observed data since dark matter fields cannot be observed directly. An immediate way to do this would be to train diffusion generative models directly on simulated galaxy catalogs with survey selection effects incorporated into the forward model for the catalogs. Certain methodological challenges may still need to be overcome: such as being able to accommodate the large number of galaxies / halos in a typical catalog. For reference, Ref. ⁵⁶ involved an application to halo point clouds with 5000 points; the SIMBIG ⁸⁷ simulated galaxy catalogs contains over 100,000 galaxies. Recent developments such as Ref. ⁹⁹ in the machine learning literature that extend the application of point cloud diffusion models to problems with over 100,000 points may provide a way out. A relevant concern that increases the *epistemic* uncertainty of this problem setting is that the highest resolution hydrodynamical simulations intended to represent DESI-like scales are often very few in number because of the significant computational cost of generating a single sample. This limits the use of generative models to regimes with enough training samples. An intriguing question to ask here could be whether work on ‘foundation models’ ^{163,29,18} could potentially provide an avenue

to extend the emulation paradigm to a “few-shot” regime. ‘Foundation models’, of which language models are by far the most successful and popular archetype, involve a learning paradigm in which by training on massive amounts of data, spanning multiple sources and domains, the model acquires the ability to perform unseen tasks. Lastly, as with all ‘black-box’ models, the use of generative models over time-tested Gaussian methods incurs the inevitable tradeoff between the promise of higher quality and “apparent realism” versus lower interpretability and risks of overfitting when learning higher-order correlations in the absence of sufficient data.

Turning to our last modality of text or code generation, Chapters 4 and 5 attempted to quantitatively probe the question “Can LLMs assist scientists in their research workflows?” in two different settings. We observed certain challenges unique to the context: in finite element analysis, the ability to manipulate a specialized finite element software proved to be a key bottleneck. In cosmology, while the LLM agent was typically familiar enough with C code to generate code that compiles, it often stumbled over subtleties in how different parts of the code are linked together. The common themes we observed included that, perhaps unsurprisingly, considerable human effort, experimentation and thoughtful planning is required to curate a framework that simultaneously permits seamless interaction with the LLMs / LLM-agents, factoring in their fast-changing capabilities and is consistent with the requirements imposed by the demarcated boundaries of the problems under focus. Relevant entangled decisions in this regard include the extent of flexibility or compartmentalization the LLM is allowed to have in the actions or changes it is permitted and the cost of experimentation, which might control whether more advanced search algorithms can be implemented. The cosmology application benefits from the existence of a “centralized” code base and a common subset

of widely recognized goals (‘tensions’) that the theoretical cosmology community has largely converged to and is working to solve. Many subfields in physics and astrophysics do not benefit from a similar centralization, since papers within the same subfield may involve considerably different research goals, with tools, datasets and resources that may not be decided apriori. Moreover, choices of datasets, code bases, and research goals, often evolve during the course of research. In this sense, the “targeted” LLM-cosmology test bed we consider here is different from the open-ended, random-walk-esque fashion in which humans often perform research. Another learning, particularly from Stage 2 of Chapter 5, pertains to the problem of *scalable oversight*³¹: even though existing LLM capabilities in the research contexts we explored do not appear to exceed those of experts in the field, identifying a generalizable set of algorithmic ‘necessary and sufficient’ tests that guarantee correctness across a wide family of problems can be challenging. Additionally, as problems become more specialized, fewer experts may be available to reliably identify subtle errors.

In the near-term, I envision the emergence of ‘LLM-research interfaces’ as a useful (and, I would cautiously hope, popular) outcome of these exercises—interfaces that seamlessly allow hybrid human-LLM collaboration. In the cosmology context, such a workflow could involve an LLM-agent navigating CLASS and making its initial pass at solving a problem, say by reading a provided paper that reports a novel cosmological theory and generating a tentative patch, even if it isn’t able to completely and correctly implement it. The next step could involve the cosmologist overseeing the task proposing changes or using their intuition to nudge the LLM to fix lingering issues. While our work in Chapter 5 adopted a minimal-interactive-human-intervention approach to document their capabilities in a ‘hands-off’ setting, our framework already permits this more interactive

workflow. Interfaces such as these could serve as a next-generation version of the popular ‘Papers with Code’ add-on to arXiv and may enable greater reproducibility, reduce the ‘equations on a blackboard’-to-code turnaround time, lower the barrier of entry for early career researchers and may hopefully, whether directly or indirectly, fuel human breakthroughs in cosmology.



Stellar-reddening-based Extinction Maps for Cosmological Applications

A.1 DETAILS OF STELLAR SELECTIONS

A.1.1 PRELIMINARY SELECTIONS

The cuts on each object require all of the following:

- $\text{nmag_ok} > 0$ in at least 2 PS1 bands and the sum of nmag_ok over all 5 PS1 bands ≥ 4 .

The nmag_ok flag is a value for each band, indicating the number of ‘okay’ detections for that band.

- $m_{PSF}^{PS1} - m_{Aperture}^{PS1} < 0.1$ in at least 2 bands. m_{PSF}^{PS1} and $m_{Aperture}^{PS1}$ denote the PSF and Aperture magnitudes for a source.
- ‘good detections’ in at least 4 out of 8 bands
- ‘good detections’ in at least 2 PS1 bands
- not extended in any of the 2MASS bands, using the ext_key flag. The 2MASS ext_key flag has a non-null value for sources whose profile is larger than the telescope PSF and were processed by the extended source processor⁵⁷.

A ‘good detection’ for each PS1 band passes the following quality cuts: nmag_ok > 0 AND the magnitude is fainter than the PS1 saturation limit AND the error is less than 0.2. A ‘good detection’ in each 2MASS band satisfies C1 AND C2, where C1 and C2 are conditions defined for each band as follows:

- $C1 \leftarrow (ph_qual='A') \text{ OR } (rd_flag=1) \text{ OR } (rd_flag=3) \text{ AND } (cc_flag=0)$
- $C2 \leftarrow (use_src=1) \text{ AND } (gal_contam=0)$

rd_flag is the ‘read flag’ that indicates the origin and quality of detections and magnitudes for a source in each band. The ph_qual flag for each band is a measure of the quality of photometry for that band, on the basis of rd_flag, scan signal to noise ratios and measurement uncertainties. cc_flag is the contamination and confusion flag. The definitions of these flags and criteria can be found in Ref.⁵⁹. The choice of PS1 and 2MASS cuts above is the same as that followed in Ref.⁸¹.

The criterion for using Gaia EDR3 parallaxes is as follows:

- the Renormalized Unit Weight Error (RUWE) must be less than 1.4
- the prediction of the astrometric fidelity classifier must exceed 0.5

A.1.2 SVM BASED CUTS

To generate a spectroscopically matched subset of objects, we query all stars at $b > 50^\circ$ from the catalog obtained after the preliminary set of selections (Section 3) and cross match to the SDSS DR17 specobj catalog. This gives us a much smaller subset of objects, all of which have spectroscopic labels (`sdss_dr17_specobj.CLASS`). The class of an object is either ‘STAR’, ‘GALAXY’ or ‘QSO’ – for both models, an object with a class of ‘STAR’ is assigned a label of 1, and 0 otherwise. We perform the following quality cuts to identify objects with reliable spectra – $rchi2 < 2$, $chi68p < 2$, and $sn_median_all > 10$.

To train the WISE SVM cut, we further identify objects with reliable magnitude measurements for the relevant features by requiring the magnitude errors in the PS1 bands r , i and z and WISE W1 band to be less than 0.2 and $w1rchi2 < 2$, yielding 158154 objects. We then generate a balanced training set consisting of an equal number of objects with labels 0 and 1. We then train an SVM with a linear kernel, that takes two features as its input: $m_r - m_i$ and $m_z - W1$, using the SCIKIT-LEARN package to learn the boundary in the left panel of Figure 1.2. When applied to an arbitrary object in the full catalog, objects with values of $z-W1$ lying above the SVM boundary were eliminated. The cut is implemented such that objects without a detection in W1 (`allwise.w1pro=0`) are assigned a ‘limiting’ magnitude $W1_{lim}=17.44$, which corresponds to

Table A.1: Distribution of objects from the spectroscopically matched sample in the Northern Galactic Hemisphere in different classes as stellar selection cuts are incrementally applied. The number reports the number of objects in each class. The figure in brackets reports the fraction of objects belonging to a certain category after a particular choice of cuts is applied $\frac{n_{Class}^{AfterCuts}}{n_{Total}^{AfterCuts}}$. Each row in the table also includes all the cuts in the rows preceding it. (The percentages in each row sum to 100%).

Cuts	Stars: No./ Proportion (%)	QSOs: No./ Proportion (%)	Galaxies: No./ Proportion (%)
No Cuts Applied	175597 (83.33)	34817 (16.52)	299 (0.14)
+ $d_{50\sin(b)} > 400pc, \sigma_\mu < 1.5$	164298 (83.04)	33292 (16.82)	272 (0.14)
+WISE nondetection/ $(r - i, z - W1)$ cut	161754 (99.01)	1436 (0.88)	178 (0.11)
+EDR3 parallax detection	161380 (99.04)	1436 (0.88)	121 (0.07)
+SDSS $(u - g, g - r)$ cut	160019 (99.84)	151 (0.09)	102 (0.06)
+ $\chi_r^2 < 3$	144750 (99.94)	42 (0.03)	49 (0.03)

the 96th percentile of objects with reliable spectra and detections in $W1$. The rationale behind this is that an object without a detection in $W1$ would be fainter than $W1_{lim}$, or that if it had been detected, $W1 > W1_{lim}$. Thus $z - W1 < z - W1_{lim}$. Since the SVM criterion for including a source is for $[m_z - m_{W1}] < b + A \times [m_r - m_i]$, if $z - W1_{lim}$ lies below the SVM boundary, so would $z - W1$.

To train the SDSS SVM based cut, we take the same set of spectroscopically labelled objects as above and perform the same cuts to select objects with reliable spectra. We further require magnitude errors in the u, g , and r bands from SDSS DR14 starsweep to be less than 0.2 and filter out objects with invalid (NaN) magnitudes or with magnitudes of exactly 22.5 in u, g , or r , yielding 203492 objects. As above, we balance the training set to consist of an equal number of objects with label 0 or 1. We then train an SVM based classifier with a radial basis function kernel that takes two features as it's input: $m_u - m_g$ and $m_g - m_r$.

Table A.2: Ablation Tests on the spectroscopically matched sample in the Northern Galactic Hemisphere: All cuts in the Secondary Stellar Selection EXCEPT a given cut are applied.

Excluded Cut	Stars: Number	QSOs: Number	Galaxies: Number
No cuts excluded	144750	42	49
-WISE nondetection/ $(r-i, z-W1)$ cut	145950	106	55
-EDR ₃ parallax detection	145065	42	60
-SDSS $(u-g, g-r)$ cut	145504	210	57
$\chi_r^2 < 3$	160019	151	102

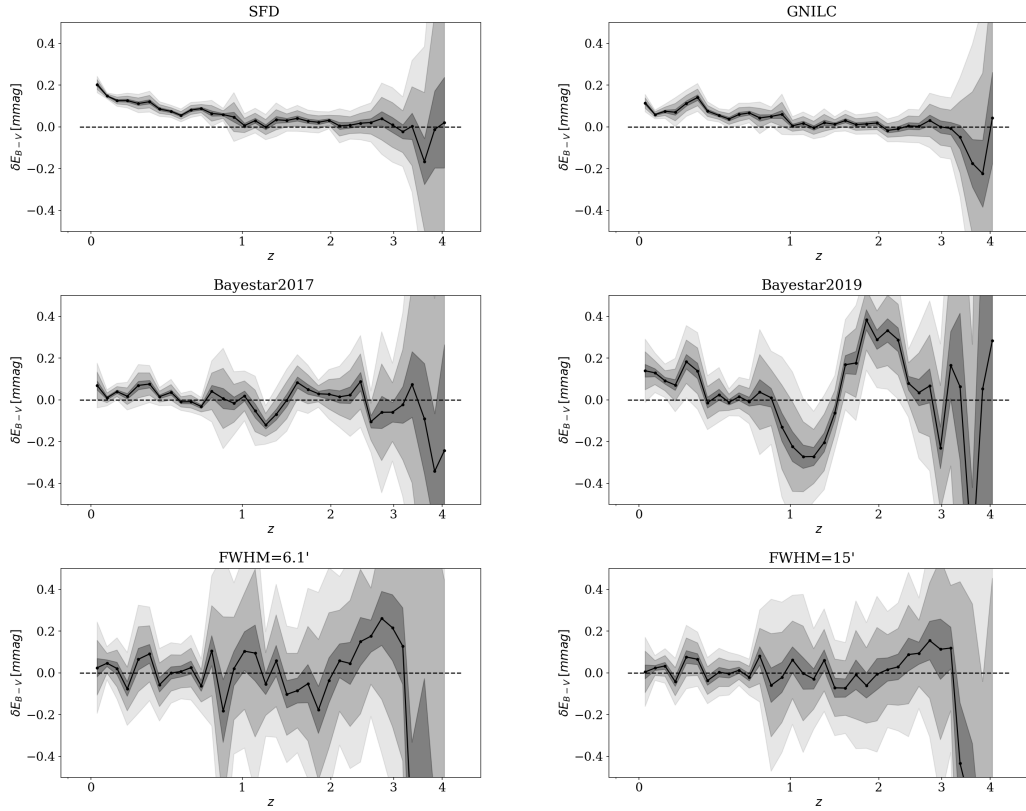


Figure A.1: Excess extinction in maps at the positions of reference extragalactic objects in redshift bins spanning $z=0-4$ for existing emission-based maps (top panel: SFD and the Planck 2016 (GNILC) dust emission map), existing stellar-reddening based maps (middle panel: Bayestar17 and Bayestar19) and our reconstructions with FWHMs of 6.1' and 15' (last panel). The sigma contours are obtained by bootstrapping over the reference sample objects at each redshift.

A.2 DERIVING THE ANGULAR CROSS CORRELATION ESTIMATOR

In this section, we describe how the angular cross correlation is computed, largely following the prescription in Ref.⁴⁵. φ indicates a position (pixel) on the sky and $E(\varphi)$ an input dust map. $\varphi + \theta$ denotes all coordinates lying at an angular separation of θ from the pixel φ . $\delta_r(\varphi, z)$ denotes the fractional overdensity of all reference extragalactic objects in redshift bin z , (denoted by $\mathcal{R}(z)$), lying in the pixel φ . Thus, $\delta_r(\varphi + \theta, z)$ is the fractional overdensity field of reference objects belonging to $\mathcal{R}(z)$ evaluated in the ring with its center at φ and radius θ . $\langle \delta E^{EG}(\theta, z) \rangle_r$ denotes the average (over all pixels in the masked region) excess extinction at an angular separation of θ around reference extragalactic objects at redshift z . $\langle \Delta E(z) \rangle$ is the integral of $\langle \delta E^{EG}(\theta, z) \rangle_r$ over all angular bins from 0 to θ_{max} divided by θ_{max} and is the δE_{B-V} quantity plotted on the y axis, as a function of z in Figures 1.4, 1.5, and A.1.

$$\delta E(\varphi) \leftarrow E(\varphi) - \text{Smooth}(E(\varphi))_{\sigma_{GAL}} \quad (\text{A.1})$$

$$\delta E(\varphi) \leftarrow \delta E(\varphi) - \langle \delta E(\varphi) \rangle \quad (\text{A.2})$$

$$\langle \delta E^{EG}(\theta, z) \rangle_r = \langle \delta_r(\varphi, z) \delta E(\varphi + \theta) \rangle_\varphi \quad (\text{A.3})$$

$$\langle \Delta E(z) \rangle = \frac{1}{\theta_{max}} \int_0^{\theta_{max}} \langle \delta E^{EG}(\theta, z) \rangle_r d\theta = \vec{W}_z^T \vec{\delta E} \quad (\text{A.4})$$

The cross-correlation weight matrix for a given redshift z is stored as $W_z(\varphi)$ where φ is the set of pixels at NSIDE=2048 over which the cross-correlation signal is calculated.

$$\begin{aligned}
\langle \Delta E(z) \rangle &= \frac{1}{\theta_{max}} \int_0^{\theta_{max}} \langle \partial E^{EG}(\theta, z) \rangle_r d\theta = \frac{1}{\theta_{max}} \int_0^{\theta_{max}} d\theta \langle \partial_r(\varphi + \theta, z) \partial E(\varphi) \rangle_\varphi \\
&= \frac{1}{\theta_{max}} \sum_{\theta_i=0}^{\theta_{max}} \Delta \theta_i \sum_{\varphi} \frac{1}{N(z) N_{pixels}} \sum_{r \in \mathcal{R}(z)} \delta^K(r, \varphi + \theta_i) \partial E(\varphi) = \vec{W}_z^T \vec{\partial E} \quad (A.5)
\end{aligned}$$

where, $W_z(\varphi) = \frac{\sum_{r \in \mathcal{R}(z)} \sum_{\theta_i=0}^{\theta_{max}} \Delta \theta_i \delta^K(r, \varphi + \theta_i)}{\theta_{max} N(z) N_{pixels}}$ and $N(z) = |\mathcal{R}(z)|$

$\delta^K(r, \varphi + \theta_i) = 1$ if reference object r lies in the ring with its center at φ and radius θ_i , 0 otherwise. N_{pixels} is the number of pixels in the mask on the map over which the correlation is evaluated and $\Delta \theta_i$ is the integration coefficient corresponding to the angular bin θ_i . Figure A.1 depicts the angular cross-correlation plots for the six maps in the first three rows of Figure 1.5 using error bars derived by bootstrapping over the reference objects. The maps we construct still possess correlations within the $1 - 3\sigma$ level. The difference between the two error propagation methods is most apparent for the emission-based maps, with GNILC now appearing to possess correlations at greater than the 5σ level.

B

Diffusion Models for Astrophysical Fields

B.1 EXPERIMENTS AND COMPUTE TIME

The 64×64 model took 1.25h to train for 60k iterations while the 128×128 models took 13.5 hours to train for the same number of iterations. We used an NVIDIA A100 for most experiments and runs. We ran most experiments on the 64×64 model, since it took 1.25 hours to train for 60k iterations. The cumulative compute time spent on experiments was around 5 days. Sampling 10 images from the trained diffusion models took ~ 40 seconds for the 64×64 model and ~ 110 seconds for the 128×128 model on the A100. We gauged whether other hyperparameter choices were better by plotting the summary statistics of samples from the generated model (as in Figure 2.2).

- We tried the cosine learning schedule proposed in ¹⁵² and found the linear schedule to work better, for the number of iterations we experimented with.
- We chose a lower learning rate for the dust images since we found this to be more stable.
- We also tried learning rates of $[6 \times 10^{-5}, 2 \times 10^{-4}, 1 \times 10^{-3}]$ for the 64x64 model and did not find any of the other learning rates to improve performance.
- We examined the summary statistics for samples from the 64x64 runs described in Section 3 for the last 15 checkpoints and did not find significant differences in the quality of the summary statistics.



Diffusion-HMC

C.1 PARAMETER INFERENCE

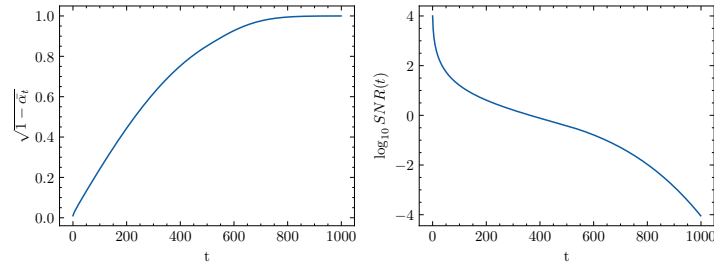


Figure C.1: The scale / standard deviation of the cumulative noise in Equation 3.1 added to the image over different timesteps.

In Equation 3.5, we set $\mathbf{M}^{-1}$ to be diagonal with $[1, 5]$. Setting the inverse mass matrix in an HMC to be close to the covariance of the expected posterior distribution helps the chain explore the space

better. In this case, we choose a step size of 5×10^{-4} , because of the steep gradient of the posterior distribution with respect to Ω_m . For the parameter on field where the true parameter is 0.101, i.e. on the prior range, we need to further reduce the step size to 1×10^{-4} in order for the parameters to be accepted. We modified the `Hamilton` package in order to generate samples. For $T_{\text{MAX}} \geq 20$ we compute the contributions to the VLB loss in batches of 10 timesteps and accumulate the gradient contribution for each batch. This enables us to compute the gradients with 1000 timesteps. The choice of mass matrix accelerates the chain’s convergence to the correct region of parameter space for σ_8 . We approximate $-\log p_\varphi(x_0|\theta)$ with the truncated variational lower bound in Equation 3.4. While we do not compute the expectation over multiple seeds *within* a single evaluation of Equation 3.5, for speed, every evaluation uses a different seed and noise pattern. The prior is chosen to be a flat prior over $\Omega_m \in [0.1, 0.5]$ and $\sigma_8 \in [0.6, 1.0]$.

$$\begin{aligned}
L_{vllb} &= L_0 + L_1 \dots L_{T-1} + L_T = & (C.1) \\
&= \mathbb{E}[-\log p_\varphi(x_0|\theta)] \leq \mathbb{E}_q[\text{D}_{\text{KL}}[q(x_T|x_0)||p(x_T)]] + \sum_{t \geq 1} \text{D}_{\text{KL}}[q(x_t|x_{t+1}, x_0)||p_\varphi(x_t|x_{t+1}, \theta)] \\
&\quad - \log p_\varphi(x_0|x_1, \theta)
\end{aligned}$$

NOTATION:

- x_0 : Normalized input field
- φ : Noise Model (Neural Network)

- θ : Conditioning cosmology i.e. a vector with Ω_m and σ_8

COMPUTING VLB TERMS:

$$L_0 = -\log p_\varphi(x_0|x_1, \theta) = -\ln \mathcal{N}(x_0|\mu_0, \beta_0) = \sum_p \frac{(x_0 - \mu_0)^2}{2\beta_0} + 0.5 \ln |2\pi\beta_0|$$

$$\text{For } t \in [1, T-1], L_t = D_{KL}[q(x_t|x_{t+1}, x_0) || p_\varphi(x_t|x_{t+1}, \theta)]$$

$$q(x_t|x_{t+1}, x_0) = \mathcal{N}(\tilde{\mu}_t(x_{t+1}, x_0), \tilde{\beta}_t)$$

$$\tilde{\mu}_t(x_{t+1}, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1 - \bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t(1 - \bar{\alpha}_{t-1})}}{1 - \bar{\alpha}_t}x_{t+1} \text{ and } \tilde{\beta}_t = \beta_t \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}$$

$$p_\varphi(x_t|x_{t+1}, \theta) = \mathcal{N}(\tilde{\mu}_t(x_{t+1}, \hat{x}_{0,\varphi}), \tilde{\beta}_t)$$

$$\hat{x}_{0,\varphi} = \frac{x_{t+1} - \sqrt{1 - \bar{\alpha}_t}\varepsilon_\varphi(x_{t+1}, t, \theta)}{\sqrt{\bar{\alpha}_t}}$$

C.1.1 HMC CONVERGENCE

For a single field, we examine the convergence of parameters, when the chain starts from different initial parameters in Figure C.2. The chains are indistinguishable beyond around 50 samples. We thus choose a burn-in of 100 samples. The $\hat{R}$ for both parameters computed using the samples in [100–300] is 0.997. A $\hat{R}$ of greater than 1.1 usually indicates that the chains have not converged and still retain some memory of their initialization. While the $\hat{R}$ is theoretically expected to be around 1 or slightly greater, some numerical variation about this expected value can result in values that are slightly less than 1. The $\hat{R}$ is a measure of the variance between chains divided by the variance within

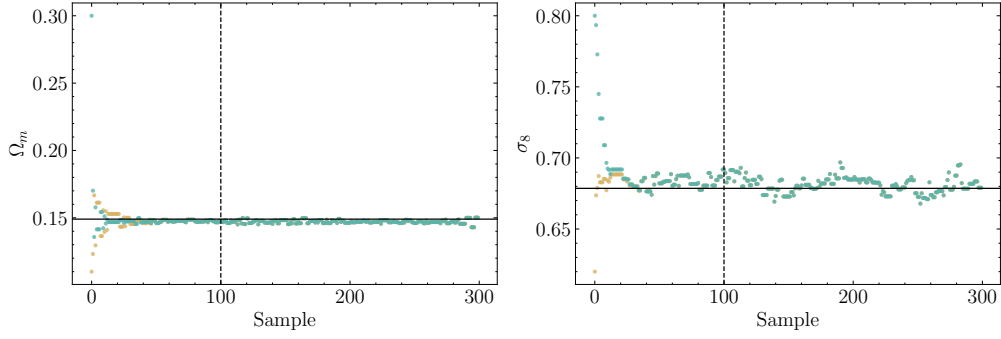


Figure C.2: HMC chains for 300 samples, for the same field, starting at two different initial parameters: $[0.11, 0.62]$ (beige) and $[0.3, 0.8]$ (teal). The chains are well mixed by the cutoff we designate as our burn-in (100 samples), denoted by the dashed line.

chains.

C.1.2 PARAMETER ESTIMATION BASELINES

Power Spectrum NPE Baseline: We use a Masked Autoregressive Flow¹⁶¹ to implement the normalizing flow that predicts the 2 dimensional parameter vector given the 129 dimensional feature vector for a single field (128 bins for the power spectrum+1 for the mean of the log fields). The power spectrum is the log of the overdensity power spectrum of the (unlogged) fields. In Figure 3.6, the power spectrum sample contours are smoothed by convolving with a Gaussian kernel with a scale of 0.8 and the Diffusion-HMC samples are smoothed by a kernel of 0.2. The ellipses in the inset figure are computed using the covariance of the 400 diffusion model samples and finding the ellipses corresponding to the 68.3, 99.4 and 99.7% confidence intervals, using the eigendecomposition of the covariance.

C.1.3 ADDITIONAL ROBUSTNESS TESTS

DROPPING THE PRIOR: In Figure C.3, we explore robustness without the prior in the HMC setting, for the noise levels in Figure 3.7 as well as with the addition of more noise. For small amounts of noise, ($\sigma = 0.01, 0.02$) removing the prior doesn't affect the bias of the Diffusion-HMC-inferred parameters since the numbers in Figure 10 are identical / close to those in Figure 3.7. For noise with a scale of 0.05, the Diffusion-HMC constraints are more perturbed for Ω_m but more robust for σ_8 relative to the neural network baseline. A standard deviation of 0.05 is equivalent to the diffusion noising timesteps between the 1st and the 2nd timestep, while $\{0.01, 0.02\}$ are less than the first timestep. Note, we plot the power spectra of the 'true' (*linear*) mass density field in these figures, to be consistent with the power spectra in Figures 1 and 2. However, the noise is added to the log of the field. Thus while the effect on the linear power spectrum is mild (around 5% at the smallest scales) for $\sigma = 0.05$ the effect on the power spectrum of the log field is around 35% at the smallest scales.

DROPPING INITIAL TIMESTEPS: In the left panel of Figure 3.5, we observed that the isolated effect of the smallest timesteps also had the strongest constraints. We thus ask the question, if we perform parameter inference in a setting where there is some knowledge of the amount of noise added, can dropping timesteps confer additional robustness? In Figure C.4, we add a comparison for the noising amounts corresponding to 0.02 and 0.05 and drop the first 2 timesteps. The rest of the setting is the same as in Figure C.3 i.e. we also drop the prior. We find that dropping these timesteps

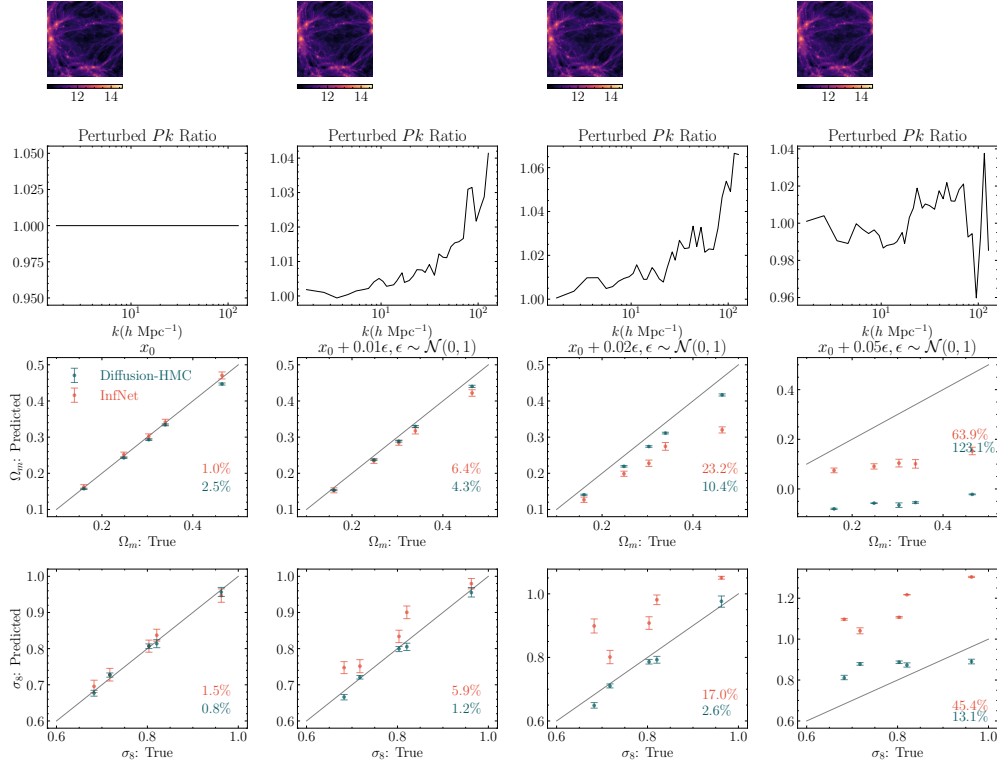


Figure C.3: Robustness comparison without a prior applied in the HMC case.

confers greater robustness on the estimates, without noticeably reducing confidence. While we defer a more rigorous investigation to future work, the prospect of using the knowledge of the amount of noise added to strategically drop timesteps could be of interest, reminiscent of scale-dependent analyses and scale cuts in other cosmological analysis methods^{61,116}.

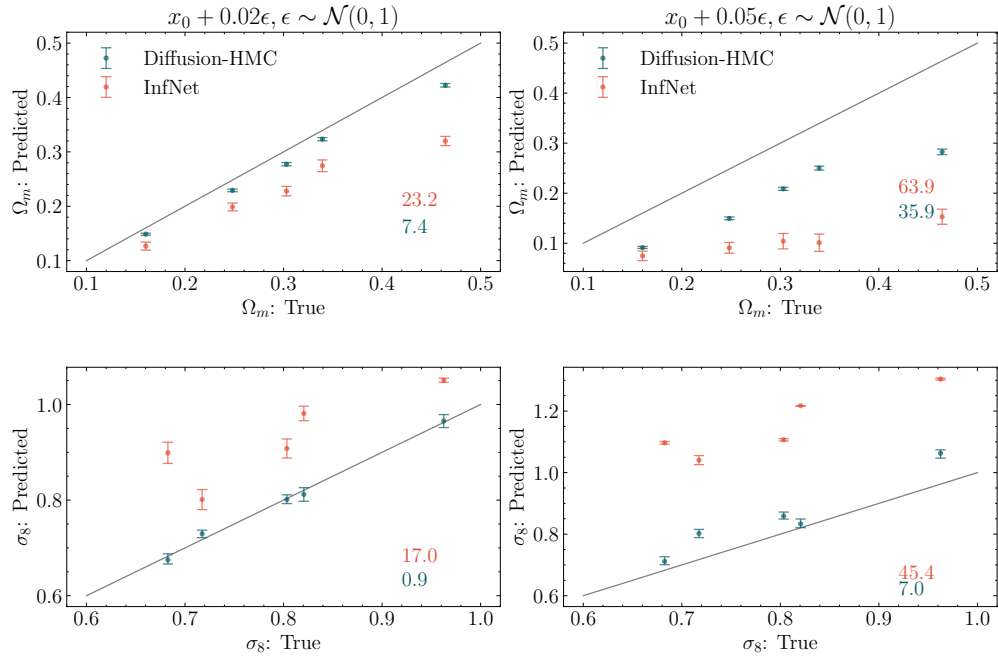


Figure C.4: Robustness comparison with the first 2 noising timesteps dropped *and* without a prior applied in the HMC case.

C.2 SUMMARY STATISTICS

C.2.1 CHI-SQUARED STATISTICS

For the CV fields, where we have 405 samples of the true and generated fields for a single parameter, we compute the chi-squared statistic using an estimate of the covariance between different k bins.

The number of k bins here is 35. In the equations below, Pk_{Ref} is the distribution of power spectra of a subset of fields from the true distribution, and is used to construct the reference mean, μ and covariance estimate $\hat{C}^{-1}$. For the power spectrum of a single field Pk_{Test} , the multidimensional chi-squared statistic is $\chi^2(Pk_{Test})$ and is computed as below. $\chi^2(Pk_{Test})$ is computed for each field in the test (generated) sample, and the mean of the distribution of chi-squared statistics is computed.

$$\vec{\mu} = \langle Pk_{Ref} \rangle \quad C = \text{Cov}[Pk_{Ref}] \quad \hat{C}^{-1} = C^{-1} \frac{N - p - 2}{N - 1} \quad (C.2)$$

$$\chi^2(Pk_{Test}) = (Pk_{Test} - \vec{\mu})^T \hat{C}^{-1} (Pk_{Test} - \vec{\mu}) \quad (C.3)$$

For the LH fields during model selection, since we just have 15 fields in the true dataset, we cannot reliably estimate a covariance. We use the following formula instead. The number of k bins here is 128.

$$\chi_r^2(s) = \frac{1}{|k|} \sum_k \frac{(P(k)_s - \langle P(k)_{TRUE} \rangle)^2}{\sigma[P(k)_{TRUE}]^2} \quad (C.4)$$

C.2.2 ACROSS CHECKPOINTS

For the eight checkpoints we generated 500 fields, with 50 fields for each of the 10 validation parameters. We then examined how different the reduced chi-squared statistic of the power spectra of the log fields, the linear fields and the p values of the means of the distributions of true and generated fields. A value of less than 0.05 would indicate that the two distributions of the means are statistically different. The p values are above 0.05 for all eight checkpoints. Since these comparisons are limited by the number of samples in the true set for each parameter (15), we additionally plot the reduced chi-squared statistic derived by using each true fields as the test and the other 14 fields as the reference. While there is some oscillation across checkpoints for each of these three statistics, the variation appears random.

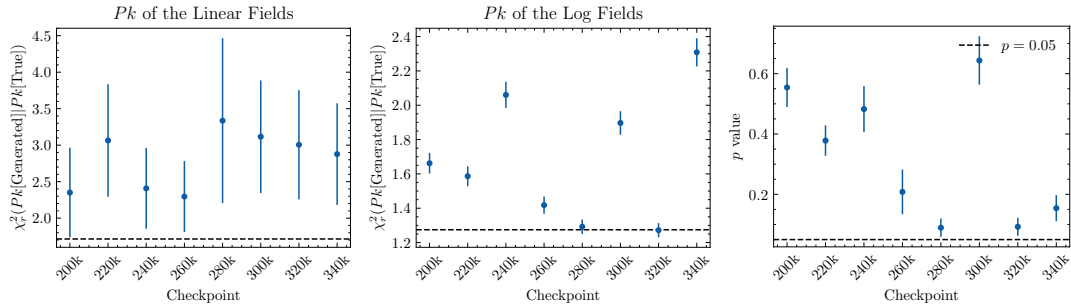


Figure C.5: Mean and standard error on the mean for the reduced χ_r^2 statistic of the power spectra of the 50 linear (left) and log (center) fields for each parameter relative to the 15 true fields' power spectra for that parameter, across 10 different parameters for each of 8 checkpoints. The dashed line demarcates the mean reduced χ_r^2 of the true fields using the other 14 true fields as the reference distribution (leave-one-out). **Right:** Mean and standard error on the mean of the p values of the distribution of the means of the 50 generated log fields relative to the distribution of the means of the 15 true fields for the same parameter. The p values are above 0.05 for all of the 8 checkpoints.

D

FEABench: Evaluating Language Models on Multiphysics Reasoning Ability

D.1 RELATED WORK

LLMs AND AGENTS FOR CODE Several studies have focused on benchmarking coding in general-purpose programming languages, with a particular focus on software engineering tasks^{19,41,106,129}, and less commonly, science problems²²². FEA software emerged because simulating and numerically solving real-world problems from scratch in mainstream languages would require significantly more effort without specialized packages, automatic mesh generation and pre-verified physics mod-

ules. Other work in the LLM literature has focused on optimizing agent-tool call and design such as the ReAct and CodeAct strategies^{234,243}. It would be valuable to port blocks from our agent such as the Evaluator and the specialized functions into generalist agentic frameworks like AutoGPT and LangChain^{210,39} to explore possible performance gains and understand the optimum way to distil visual information from the Graphical User Interface (GUI). Beyond the realm of general-purpose programming, some works have sought to incorporate productivity APIs such as those for weather, email among others into agentic workflows^{175,23}. Our agentic approach shares similarities with the Reflexion strategy²⁰⁸, although in our case the Evaluator mainly subjective feedback from the API, and only queries its VerifierLLM when executability is already high.

LLMS FOR SCIENCE The utility of LLMs in science has been explored by evaluating their performance on tasks in medicine^{195,242}, theorem proving²⁴¹, examination problems of varying levels of difficulty^{93,233,127} and in specific domains such as physics and chemistry^{159,32}. More recently, there have been efforts to examine whether LLMs can be of utility in other aspects of the scientific process, such as developing hypotheses, reproducibility of code and question-answering^{172,141,209}. Ref.¹⁵¹ and Ref.²²¹ made a preliminary exploration on getting LLMs to solve elasticity problems and in a human-in-the-loop setting and Ref.¹¹⁸ explored the role of LLMs on optimizing airfoils.

D.2 DATASET CURATION

D.2.1 FEABENCH GOLD

SELECTION CRITERIA:

We chose tutorials that satisfied the following considerations:

1. *Simpler Geometry*: COMSOL Multiphysics® can be used to analyze the physics of systems involving intricate geometries such as microwaves or transformers. In these cases, in practice, most problems involve importing a pre-built geometry object that might have been built externally using Computer-Aided Design (CAD) software and to then perform the remaining analysis. Since we wanted to explore the ability to solve the problem end-to-end and without requiring imports of derived objects, we restrict ourselves to problems that did not require imports of geometry, or any other files.
2. *Tutorial / Code Simplicity*: We additionally chose problems that did not involve multiple ‘Model’ JAVA classes and restricted ourselves to tutorial documents with fewer than 20 pages. The first requirement is a consequence of how our connection to the COMSOL Multiphysics® sandbox is set up, and to make the problem easier for the models to attempt to solve. We additionally ensured that the problems were amenable to computing a numerical artifact.
3. *Solving Speed*: We also excluded any problems whose ground truth code took over a minute

to solve.

GENERATION PROCEDURE:

Without any modification, the tutorials might export a single value, a table, or not export any target quantity at all, with the final output being qualitative in nature, such as in the form of plots or figures. For our benchmark, however, we specifically wanted every problem to have a numerically verifiable target value, in order for there to be an absolute notion of correctness (i.e. if the code was fully correct, and aligned with the intent of the problem, it should be able to export this value). This also enables easier evaluation of the problems. The following procedure and guidelines were adopted to curate the benchmark:

- For an initial set of 2-3 problems, model specifications and plans were annotated by hand, by an expert user of COMSOL Multiphysics®.
- For subsequent problems, we speed up the benchmark generation procedure by following an initial LLM-assisted data generation process, with the final verification steps involving humans. An LLM is provided with a tutorial, as well as a two-shot prompt with the expert annotated model specifications.
- The LLM is entasked with returning a model specification for the tutorial that has the same format. This requires the LLM to identify an appropriate target value from the tutorial which it does from either the text or the figures, and returning a model specification for computing this target value.

- The LLM is then asked to create a plan corresponding to the model specifications, using a two-shot prompt with two plans. The utility of the tutorials are that the plan is closest to the GUI instructions listed in the tutorial, while model specifications is more concise.
- A ground truth code that can compute the correct value is then generated for the problem. We manually verify that the ground code when run, exports the desired target value. This step also involves simultaneously ensuring that all information required to build the model is contained in the plan, and in the model specifications by editing the LLM-generated drafts and ensuring that no Translation Errors are encountered when parsing and executing the ground truth code in COMSOL Multiphysics® using the bridge described in Appendix D.4.1 or that any errors if encountered are in non-crucial lines and do not prevent the solution from being computed. Any missing or incorrect information is fixed, and the selection_information field, that contains numerical identities of boundaries and points is also created.
- We add an instruction to export the output to OUTPUT_PATH/output.txt in the model specifications and plan.

D.2.2 FEABENCH LARGE

The input field in FEABench Large is the ‘Modeling Instructions’ section of the tutorial. The output field is the code in the first run function of the exported Java file of the built COMSOL Multiphysics® model with the following postprocessing steps applied: we append to the last line

of each ‘study’ code block in the model with a `model.study("study_tag").run()`; where “study_tag” will typically be “std1” or “std2”, and remove the block of ‘solver’ code. While the choice of including the code only in the first run function might make the mapping between instructions and lines of code less one to one in problems with more than one run function, this choice makes this dataset and the style of code resemble the constraints in FEABench Gold. We make the ‘study / solver’ changes because the ‘model.sol’ code consists of a larger block of automatically populated lines that bear little resemblance to the original problem specification, and often correspond to a single ‘Compute’ step in the GUI. Adding the ‘.run();’ line prompts COMSOL Multiphysics® to use its default solver best configured to solve the problem depending on the physics and nature of the analysis performed. This is also a pattern guiding our prompt design across tasks. The prompt used for this experiment is similar to the **Plan** One-Shot prompt.

D.3 EVALUATION DETAILS

D.3.1 BASELINE EVALUATION METRIC: CODE SIMILARITY SCORE

A **Code Similarity** score was also measured for each solution. This is a text-based similarity score between the solution and the GT code. We report this metric as a baseline measure of code similarity, and to further motivate our introduction of domain-specific metrics.

We used the `diffli`¹⁷⁴ package to compute a score between 0 and 1 as a measure of string similarity, using the ratio of the lengths of the longest matched subsequences to the ratio of the lengths of strings being compared. Code Similarity reflects this score between the generated code and the

Table D.1: Code Similarity across LLMs.

Model	FEABench Gold: ModelSpecs	FEABench Large
Claude-3.5-Sonnet	0.19 ± 0.03	0.20 ± 0.01
GPT-4o	0.17 ± 0.03	0.15 ± 0.01
Gemini-1.5-Pro	0.17 ± 0.03	0.15 ± 0.01
Gemma-2-27B-IT	0.15 ± 0.02	
Gemma-2-9B-IT	0.11 ± 0.02	
CodeGemma-7B-IT	0.12 ± 0.02	

Table D.2: Code Similarity across experiments.

Experiment	Code Similarity
ModelSpecs : One-Shot	0.17 ± 0.03
ModelSpecs : PhyDoc In-Context	0.15 ± 0.02
ModelSpecs : Multi-Turn Agent	0.17 ± 0.03
Plan : One-Shot	0.21 ± 0.03
Plan : PhyDoc In-Context	0.20 ± 0.02

ground truth code. It is not surprising that this metric has the least variation across experiments and models since significantly different blocks of code might yield the same answer. The preponderance of boilerplate syntax, along with the fact that two different code blocks could generate equivalent model subtrees, are factors that contribute to the lack of meaningful variation of this metric across experiments. As a specific example, a `model.study("std1").run();` will leverage COMSOL Multiphysics®'s default numerical solver for the problem. However, this could also be represented explicitly using large blocks of `model.sol("sol1")...` lines in the Ground Truth Code field.

D.3.2 EXECUTABILITY

The LLM output is first parsed to identify the block with Java API calls, and further parsed to pythonize the lines (Appendix D.4.1). This filters out lines that are not code or cannot be pythonized and results in a sequence of COMSOL Multiphysics® API calls and their ‘pythonized’ counterparts, all of which start with `model.` and end with `;`.

The pythonized lines are then passed to the MPh client, and replies for each line are received. We parse API replies using the following patterns. A reply containing any of the following [‘Messages’, ‘has no attribute’, ‘No matching overloads’, ‘invalid syntax’, ‘Exception’, ‘is not defined’] are considered **Syntax Errors**. Replies with [‘Ambiguous’, ‘comma’, ‘No Model set’] are **Translation errors**.

The last category category is rare in our experiments and are occasionally encountered when we tested adding new problems to the benchmark that contained lines that weren’t translated correctly in the query: the first two flag errors in the query to COMSOL Multiphysics® via Mph, while the last indicates that an action is being done on a non-existent model, which is inconsistent with the setup of the code. All other replies are designated **Correct**.

$$Executability = \frac{\text{CorrectLines}}{\text{TotalParsedLines}} \quad (\text{D.1})$$

D.3.3 MODEL TREE SCORE

The model tree representation of the model built by the language model can be extracted, and one can use the same similarity score as above to compute a similarity score relative to the target tree. We

expect this to be a more reliable measure of alignment since different blocks of code that build the same model will have the same model tree (addressing the case described in Code Similarity). Using the formula below, the score will be 1.0 if the trees are identical, and 0.0 if the trees are equivalent to a tree before any code is run.

$$ModelTreeScore = \frac{Score(LM, GT) - Score(Empty, GT)}{1.0 - Score(Empty, GT)} \quad (D.2)$$

D.3.4 VALID TARGET

There are various ways in which computing the correct value and exporting it to a table may fail:

a) the LLM's code forgets the export command to the API and no table is exported b) an empty table is exported or, c) a table containing an incorrect value is exported, such as a default value or the wrong quantity (eg: time instead of temperature). Failure modes b) and c) are far more common than a) and occur when the code is not fully correct and the partially constructed COMSOL Multiphysics® model exports nothing or an incorrect value. For instance, a partially solved model that was asked to compute the temperature at time=190s might export a table where the last value was 190 but because of errors in model construction, no temperature was exported. In such a case if the ground truth answer is say, 185°C, without verifying the physical quantity, one would mistakenly evaluate the algorithmically parsed figure 190 to be quite close to the target. In other cases, the software might export a default such as 293.15 K if the solver did not solve correctly.

If a table containing the target quantity is exported, it is first read and parsed. The last value in the table is algorithmically extracted. To address this problem, we ask an LLM (Gemini-1.5-Pro), to

extract the exported value and units from the table, if it is a match for the target description, and minimize the chances of incorrectly evaluating these failure modes as valid solutions.

Figure D.1: Prompt used to ask the LLM to evaluate the problem

You are provided with a table that was exported by a model built in COMSOL. The table * should * contain the EXPECTED TARGET QUANTITY. The following failure modes may occur when the model is not built correctly:

1. The table might be empty or might export a physical quantity that is different from the expected target quantity.
2. The table might export the same physical quantity, but the quantity is just an initial or boundary condition, or a default value that was exported, instead of the result of genuinely numerically solving the problem. You can find numbers already in the problem description in 'PROBLEM'. Default values include 20degreesCelsius, 293.15 K, 0 etc.

Carefully examine the 'TABLE' and compare it with the units and description of the expected target quantity and the numbers in 'PROBLEM' to assess whether the table exported a value that was the result of genuinely numerically solving the problem. You must return TARGET VALUE and TARGET UNITS in json format if the table was the result of genuinely solving the model, computing a solution and exporting it. Return 'N/A' for both fields if the table suffers from either of the failure modes described above.

PROBLEM: `{{ problem_description }}`

Figure D.1 (Continued)

EXPECTED TARGET QUANTITY: `{{ target_description }}`

TABLE: `{{ table }}`

REPLY:

We then compute the number of problems for which the LM was able to parse the reply and convert it to a JSON. This fraction is the number we report as Valid Target.

D.3.5 RELATIVE ERROR | STRICT

Our strict filter for whether a model has truly solved the problem is to take the subset of problems for which the problem was judged to be a valid export by the LLM, and to consider the algorithmically parsed last value. We then compute the relative error of this value against the ground truth target value. If this value is less than 10%, we consider it valid.

D.3.6 PHYSICS METRICS

The interface lines are parsed from the ground truth code by finding lines that fit the regex pattern for interface creation. Likewise for the feature creation and feature property modification lines.

Each of these lines of codes can be considered as an “Action” consisting of an Action Type (eg: Create Interface) with corresponding Arguments (eg: Interface tag, Name of the Interface, Geometry).

Create Interface: `model.component("comp1").physics().create("Interface_tag",
"InterfaceName", "Geometry_tag");`

Eg: `model.component("comp1").physics().create("ht", "HeatTransfer",
"geom1");`

Create Feature: `model.component("comp1").physics("Interface_tag").create("Feature_tag", "FeatureName", Dimension);`

Eg: `model.component("comp1").physics("ht").create("temp1",
"TemperatureBoundary", 1);`

Modify Feature Property: `model.component("comp1").physics("Interface_tag").feature("Feature_tag").set("Param", "Value");`

Eg: `model.component("comp1").physics("ht").feature("temp1").set("T0",
"1000[degC]");`

INTERFACE FACTUALITY

We check whether the Interface name exists in a list of known COMSOL Multiphysics® interfaces.

If it exists in this list, we assign it a factuality of 1, else 0.

INTERFACE RECALL

How many GT interface creation actions (ignoring Interface_tag) were also in the LM code? This checks whether the same interface was defined on the same geometry. 'nan' if there are no interfaces in the GT (not encountered in our dataset).

FEATURE RECALL

Since multiple features may be created under the same interface (eg: 2 Boundary Conditions with different temperatures), we compute the occurrences of *each* GT feature name in the GT code and in the LM code, and a recall for each GT feature name, and then average over all GT features. In our implementation, if no GT features are defined, a) AND no LM features are defined the recall is 1, b) but LM features are defined, the recall is 0.

FEATURE DIMENSION

Let F_c be all the GT features that are also created by the LM solution. Let Dim_c be the set of F_c such that the LM feature has the same dimension as the GT feature. Feature Dimension = $\frac{|Dim_c|}{|F_c|}$

This is a correctness and physics reasoning metric as opposed to an alignment-focused metric since

creating a TemperatureBoundary with dimension 2 attempts to create a 2D temperature boundary condition. Creating a TemperatureBoundary with dimension 1 attempts to create a temperature on an edge. Thus this measures the LM’s ability to correctly deduce the spatial dimension of boundary conditions or other features from the context of the problem.

FEATURE PROPERTY RECALL

This compares the modify feature property actions. It computes how many GT modify feature property actions were also in the ground truth, *ignoring* differences in Interface_tag and Feature_tag. If no GT properties are modified, a) AND no LM features are modified the recall is 1, b) but LM features are modified, the recall is 0.

D.4 QUERYING THE COMSOL MULTIPHYSICS® API FROM PYTHON

D.4.1 THE PYTHON-COMSOL MULTIPHYSICS® BRIDGE

The raw output of the LLM is a string containing COMSOL Multiphysics® API commands in Java. An interface between Python and COMSOL Multiphysics® is needed to execute this code and interact in other ways with the API. We use the Python package MPh⁹⁴ and Rpyc for this. MPh is a scripting interface built on JPyype¹ that enables a Python program to communicate with and build a model in COMSOL Multiphysics®. Each Java API command in the LM’s output can be ‘pythonized’ algorithmically. In most cases, the pythonized line is near identical to the Java line. However, due to differences in Java and Python syntax there exist some corner cases that need to

be handled separately. Eg: 'new String[]' is exclusively a Java construction, while the notation for booleans in Python is True / False as opposed to true / false in Java. Thus a 'pythonizer' is constructed that parses and translates Java API calls to their Python counterparts.

The setup involves the following assumptions: an MPh client object is created. This behaves like a stateful 'sandbox', where models can be built by LLMs, code can be evaluated, or information such as the current state of the model tree, properties under a node and the exported table can be queried and retrieved. Although multiple models can be created and set under the client, for simplicity we work with settings that involve a single model. Before running a new solution, the existing model is deleted and a new blank model is created. The LLM actions will modify this blank model. Thus, by design, all lines of code the LLM outputs, should start with 'model.' and end with ';'.

D.4.2 COMSOL MULTIPHYSICS® CODE STRUCTURE

1. *Geometry*, if any: This involves identifying the dimensionality of the problem, and constructing a representation of the object being modelled, say a cup, by creating and composing primitive shapes such as ellipses or rectangles to build the object. While already constructed geometries can also be imported from other software such as CAD, in our benchmark, we currently restrict ourselves to models for which we construct the geometry from scratch in COMSOL Multiphysics®. This typically starts with a `model.component('`comp1') .geom pattern.`
2. *Physics*: This will include specifying all the physical conditions for the problem, including

initial or boundary conditions, forces, properties or in the case of mathematics problems, the differential equation. This typically starts with a `model.component('`comp1`).physics` pattern. Some problems may additionally have lines that begin with `model.component('`comp`').multiphysics`, and set up the coupling between different kinds of physical phenomena. We categorize these lines, if any as ‘physics’ in Figure 4.3 and D.2.

3. *Material*: Creating materials and assigning them to domains. One can either assign known materials such as ‘Copper’ and the object will inherit the default properties of that material, or define a blank material and its properties such as conductivity from scratch. This typically starts with a `model.component('`comp1`).material` pattern.
4. *Mesh*: Usually a shorter step that involves meshing the surfaces of the geometry to set up elements. This typically starts with a `model.component('`comp1`).mesh` pattern.
5. *Study/Solver*: This involves specifying the conditions of the analysis and solver, such as the number of timesteps. While the solver code can be modified to override defaults, COMSOL Multiphysics® also has the ability to automatically populate the model with the default solvers most apt for a given problem. This typically starts with a `model.study` or `model.sol` pattern respectively. In Figure 4.3 and D.2, we categorize both patterns as ‘solver’.
6. *Results*: Once the numerical solver has completed the analysis, one will likely postpro-

cess the problem, in order to generate desired plots or tables. This typically starts with a `model.result` pattern.

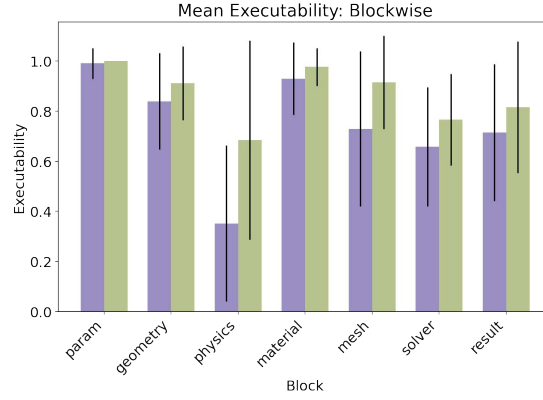


Figure D.2: Block-wise executability across the 300 initial samples of code (purple) with PhyDoc In-Context and in the best solution (green) across all problems. Error bars denote standard deviations.

D.5 AGENT DETAILS

We design a multi-agent system that interacts with the COMSOL Multiphysics[®] API, as well as tools (or specialized functions). Each agent has a specific role and input context.

CONTROLLERAGENT : The main agent that tries to solve the problem description by generating solutions, interacting with the API and calling subagents.

Input Context: Problem description.

Components: Evaluator, ControllerSubAgent

Working: This samples an initial population of $N(=20)$ solutions using PhyDoc In-Context.

Over the course of its trajectory, the agent proposes 40 solutions: 20 from oversampling the initial

prompt, and another 20 from correcting the best of the initial 20, and the best solution is selected from the tracked best solutions. This allows us to include gains obtained both from oversampling as well as from correction. For 5 problems, the best solution corresponded to one of the initial population of solutions. Each solution is evaluated by the Evaluator. A fitness score, between 0 and 2, is computed for each solution, using the following formula: $\text{Executability} + \text{ExportSuccessful}$ where ExportSuccessful is 1 if (the solution computed a value AND had executability above 90%) and 0 if not. The controller agent tracks a set of best replies using their fitness. The set of best replies stores at least $B(=1)$ solution, as well as all solutions that successfully computed a value. This agent also determines the context to be sent to the CorrectorSubAgent, using the following algorithm:

- Solution to iterate on: We use an iteration criterion inspired by the Markov Chain Monte Carlo (MCMC) acceptance criterion. The solution to iterate on (rendered in the prompt to the CorrectorSubAgent as “CURRENT CODE”) is (a) the last solution if the last solution has equal fitness as the best solution, and (b) the last solution if a random float between $[0, 1]$ is less than $\alpha = \frac{\text{Last_Fitness}}{\text{Best_Fitness}}$, else the best solution.
- ExecutionHistory: The best solutions, if not already used in context upto a maximum of 3 best solutions, in addition to the last $N_{\text{bad}}(=1)$ replies, if not already in context.

EVALUATOR : This returns the feedback for a solution in a ‘score’ dictionary (Left panel, Figure 4.2)

Input Context: An LLM solution.

Working: The evaluator always returns execution feedback and additionally includes subjective

feedback from a VerifierLLM if Executability exceeds 90%. Note, this evaluator is *not* aware of the GT target value.

CORRECTORSUBAGENT : This returns an updated solution.

Input Context: Problem description, Current Code and Feedback, Execution History

Components: ToolLookupAgent

Working: This calls the ToolLookupAgent and retrieves its reply. It then includes this reply to the rest of the context received from the ControllerAgent to propose the next solution.

TOOLLOOKUPAGENT: This calls tools and returns the information retrieved from them.

Input Context: Feedback

Components: ToolRegistry

Working: The LLM is shown tool descriptions and the input context and must return a list of tool calls, as structured classes using the Langfun¹⁶⁷ package consisting of the tool name and its arguments. If successfully parsed, each tool is called with its arguments and the replies are concatenated (see Figure 4.2 for the feedback and reply for a single step). The tools in the registry are:

1. **QueryPhysicsInterfaces:** This returns a list of valid physics interfaces.
2. **QueryPhysicsFeatures:** This returns the features under an argument *interface* or a list of known features under interfaces.
3. **QueryModelTreeProperties:** The LLM must call this tool with a *path* argument ('/physics/Heat Transfer in Solids/Solid 1' in Figure 4.2) to receive the properties under the

node corresponding to path.

4. `RetrieveAnnotatedSnippets`: To call this tool, the LLM must specify a *branch* – one of the conceptual blocks such as physics or geometry – and a *query* – a brief natural language description of a specific step. In Figure 4.2, the LLM first called this tool with the branch ‘geometry’ and the query ‘Create a 2D axisymmetric geometry in...’. A retriever then looks up the annotated library and retrieves 3 annotations along with their code snippets, most similar to the query made. Thus, this allows the LLM to search a library of code snippets to find the correct ways to express certain steps in code, simulating how a human unfamiliar with a coding language would look up similar examples of code.

At the end of this experiment, the `ControllerAgent` saves its best solutions as well as other intermediate states. During evaluation, the best solutions are read in and evaluated. If there are multiple best solutions (in cases where multiple solutions were able to compute a target value), the top best solution is the one that maximizes the following formula: $\text{Executability} + \text{bool}(\text{Computed Value}) + [(1.0 - \text{Target Relative Error}) \text{ if } (\text{Target Relative Error} < 1) \text{ AND } (\text{Valid Target}) \text{ else } 0]$. The three conditions together prioritize solutions that (1) had high executability, (2) were complete enough to export any value, albeit incorrect or the wrong quantity and, (3) exported a ‘Valid Target’ within 100% of the desired value.

The agent experiment on a single problem takes slightly over 12 minutes (ranging from 7-17 minutes) on average per problem. The dominant factor contributing to this variability is the number of LLM queries: in problems where executability crosses 0.90, there will be more LLM queries since

the Evaluator additionally calls the VerifierLLM. The FEA runtime is only a small fraction of this time: parsing the LLM reply, evaluating it by executing it in COMSOL Multiphysics® and retrieving API messages took around 0.9-1.5s for a single LLM reply. We used a subset of 5 problems to compute these estimates.

D.5.1 TOOLS

In our implementation of the ToolLookupAgent, if the tool call fails, the ToolLookupAgent will return an empty reply. Tool calls fail when the LLM is unable to generate a call that is formatted in the way Langfun expects.

QUERYMODELTREEPROPERTIES

In order to help the LLM learn how to appropriately format a valid path, say to the ‘Solid’ feature, the current state of the model tree is shown to the ToolLookupAgent LLM. It also has a history of unsuccessful (incorrectly formatted) paths in previous queries to this tool, in order to minimize the chances of incorrectly calling this tool with an invalid path.

RETRIEVEANNOTATEDSNIPPETS

We use the Discovery Engine API⁷⁷ with the model name ‘semantic-ranker-512-003’ to rank and retrieve the top 3 annotations most similar to the query snippet. The annotation library was generated by taking tutorials and splitting them into code blocks using the patterns described in D.4.2.

There are 768 pairs of annotations and snippets across all branches of code. Here is an example of an

annotation ‘summary’:

Summary: Defining a transient study with a time range from 0 to 0.025 seconds with a step of 1 second. The study will solve for the “spf” physics interface, and a relative tolerance of 0.001 will be used. The number of solver iterations will be automatically determined based on the time step.

D.5.2 ANALYSIS

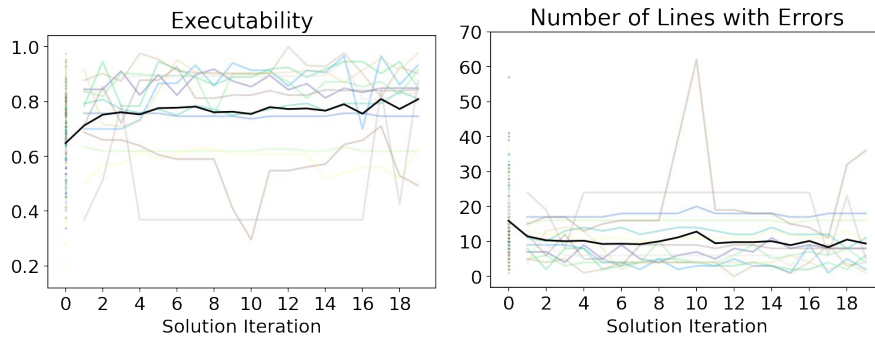


Figure D.3: Executability and number of errors over solutions returned by the ControllerAgent. The scatter at 0 denotes the spread from the samples in the initial population and the black line denotes the mean value for the metric at that state across all problems. Each colored line demarcates a different problem.

Figure D.2 depicts the blockwise executability in the initial sample relative to the best solution across problems. The standard deviations in the best case are higher since we have 1 best solution for each problem, and 20 samples per problem in the initial population. Figure D.3 plots the Executability as well as the number of errors over solution iteration. The evolution of the metrics isn’t monotonic and in some cases the agent gets stuck on the same solution for some iterations, or takes an incorrect turn. We added the acceptance criterion to minimize the number of iterations required to “escape” an incorrect turn.

D.6 QUALITATIVE ANALYSIS

We qualitatively compare the LLM-generated code for the **ModelSpecs** task in the baseline (one-shot) setting with Gemini-1.5-Pro, relative to the ground truth code field. At a high-level, the LLM’s solution consists of API calls that qualitatively possess the same structure and grammar as in the GT code. This problem requires the LLM to represent the cylindrical cross-section of a cylinder as a rectangle in 2D, with the axisymmetric condition applied for rotational symmetry about the cylinder’s axis. The LLM instead creates a 3D geometry and attempts to create a rectangle. This doesn’t work as is indicated by the error message, since the rectangle is a 2D construct and cannot be directly created as a 3D object. Since the rectangle creation action fails, no ‘r1’ node is created, and subsequent actions acting on the ‘r1’ node are invalid. This pattern of non-executability is also observed downstream, where all actions on the ‘ht’ node are rendered invalid because the ‘ht’ node could not be created in the first place. Note, if the LLM had chosen a 2D axisymmetric geometry, the remaining geometry lines of code would be correct. They fail because of an incorrect decision made first.

The LLM tries to create a ‘HeatTransferInSolids’ interface. We described this pathology in 4.5.1. In this example, the Interface Recall and Feature Recall metrics are 0, as is the Interface Factuality metric. The GT code modifies 5 features, of which the LLM only modifies 1 (setting T_0 to 273.15 K). Thus the Modify Feature Property score is 0.2.

D.7 CAN LLMs SOLVE THESE PROBLEMS IN PYTHON?

Our experiments underscored the challenge of composing physics reasoning skills with the ability to generate syntactically correct and consistent COMSOL Multiphysics® API code. In this section, we seek to evaluate how well LLMs can solve FEABench Gold problems with a general purpose programming language, Python. We set up a Docker environment with the numerical packages, `numpy`⁸⁹ and `scipy`²³², along with the open-source FEniCSx (`dolfinx`)²² package. We further test the SWE-Agent²⁴⁰ framework with `claude-3-5-sonnet-20241022` as the underlying LLM on its ability to solve these problems, since `claude-3.5-sonnet` generally performed best in Table 4.2 and 4.3.

In this setting, the problem statement for each problem includes the **Model Specifications** and the **Target Description** fields. As in the experiments in COMSOL Multiphysics®, the LLM needs to generate the code that solves the problem, and export the target value and the target units to an ‘output.txt’ file. We then apply the patch containing the code and the output, and evaluate the output. We used a budget limit of \$1 per problem. Note, in this setup, we evaluate the last ‘output.txt’ file. If the agent’s attempt at solving the problem terminates because it exceeds the cost limit, the ‘output.txt’ file will correspond to the last code that the agent successfully executed solution in the loop, and *not* the last version of the code.

Although the environment contains FEniCSx, which is likely better suited to solving FEA problems, we explicitly instructed the agent to use only `numpy` and `scipy`. We made this choice since preliminary tests revealed that the agent also failed to generate executable FEniCSx code due to breaking

changes in the package.

We evaluate the following metrics:

- **Valid Target:** As in Section 4.4, this is the number of problems for which the LLM judged the computed output to be valid i.e. consistent with the target description and units. $11/15$ solutions had a valid target.
- **Relative Error| Strict:** Only 4 solutions passed the strict cut, i.e. computed a valid target AND a relative error of less than 10%. The mean relative error over this subset is $2.15\% \pm 1.89\%$.

Over the larger set of 11 problems for which any solution was computed, the mean relative error is $296\% \pm 200\%$, where the error bars denote the standard error on the mean.

In Python, and particularly with popular ‘in-distribution’ packages like numpy and scipy, the difficulty of generating syntactically correct code is minimal. However, now the LLM must define and implement the equations describing the physics and the mesh from scratch and configure its own solver settings as it can no longer import pre-verified physics modules and automatic solvers as in the engineering simulation software. Thus the bottleneck has now shifted from generating executable code, to being able to achieve the desired numerical precision, since in only 4 cases the LLM was able to compute a value within 10% of the desired target. The emergence of commercial softwares like COMSOL Multiphysics[®], Ansys[®]¹² and Abaqus[®]FEA⁶² arose from the challenge of solving complex engineering problems with general purpose languages, and to fulfil the need for already validated physics modules, numerical precision and automatic solvers.

E

An LLM-driven framework for cosmological model-building and exploration

E.1 COSMOLOGICAL MODELS

E.1.1 PROBLEM SPECIFICATION

We provide $\mathcal{P}_{Fiducial}$ as a JSON dictionary that can be used with the Python wrapper to C, Classy.

In addition to the JSON, we include an optional ‘ini’ version of $\mathcal{P}_{Fiducial}$ that may be used to com-

pute observables using the modified CLASS executable directly, in case errors were encountered with the Classy wrapper. In addition to describing the theory, and added parameters, the standard description.md also includes implementation details to align the agents assumptions on the codebase and the input parameter with that of the target implementation. The implementation details section may include information about specific modifications and / or helpful information to address typical agent errors in preliminary experimentation, such as implicit assumptions about the structure of the codebase that would work with the provided parameter file $\mathcal{P}_{Fiducial}$. This is done so that we can consistently use the same $\mathcal{P}_{Fiducial}$ dictionary with both the LLM's implementation, as well as the target repository (when we pre-save observables for comparison in Stage 2).

We additionally include a `problem_context.yaml` file with the following structure and fields.

- stage2:
 - type: This entry confers our framework with the ability to accommodate problems with and without a target / ground truth implementation. Options:

`artifact-based | none`
 - presave_in: If 'type' is artifact-based, this entry determines whether the target implementation's observables were presaved using the C executable or Classy. Options:

`Classy | C_exec`
 - presaved_observables: If 'type' is artifact-based, this entry determines which observables should be presaved using the target repository. Eg: `['mPk', 'Omega']`
 - save_observables: Which observables to save during Stage 2 of the agentic experi-

ment. Eg: ['cls', 'mPk', 'Omega', 'H0', 'sigma8', 'S8']

– model_specific_tests:

* comparison_qty: Which observables to compare between the target implementation and the agent’s implementation.

* field: Which dark energy density parameter is relevant to the problem. Eg:

$$\Omega_{scf}, \Omega_{fld}$$

• stage3:

– type: Whether to find parameters that minimize the negative log-likelihood

(minimization) or ask the agent to manually propose and explore the model’s behavior in different limits on its own (exploration).

– vary_params: Single-line instruction explaining which parameters need to be minimized / explored.

• extra_context_files: This allows us the flexibility to include additional problem-specific contextual prompts for an experiment. Eg: ["propose_model_prompt.md"] for *Hypothesis Generation*

E.1.2 PROBLEM INSTANCES

EARLY DARK ENERGY

The dark energy density parameter of interest here is Ω_{scf} .

THAWING QUINTESSENCE

The agent is instructed to minimize Ω_{cdm}, h and the added parameters. The density parameter of interest here is Ω_{fld} . The `description.md` we used is in Figure E.1. The body includes the following sections: the description of the theoretical framework, $\mathcal{M}$, and required equations, the added parameters that describe the model, any required background field variables and implementation details: specific cues that assist the LLM in making the desired changes to CLASS.

Figure E.1: Description for the *Thawing Quintessence* model.

The Thawing Quintessence (Tanh Parametrization) Model

This is a thawing quintessence model characterized by a specific parametrization of the dark energy equation of state, $w(z)$. Thawing quintessence models propose that the dark energy is a scalar field that was initially frozen (mimicking a cosmological constant with $w \approx -1$) due to Hubble friction at high redshifts. As the universe expands and Hubble friction decreases, ϕ eventually unfreezes. At that point, w departs from -1 , and the ensuing evolution of ϕ and w depends on the scalar potential $V(\phi)$. the field begins to "thaw" and evolve, causing w to deviate from -1 .

The core of this model is a phenomenological parametrization for the equation of state $w(z)$ designed to capture the qualitative behavior of thawing quintessence in a model-independent way:

$$w(z) = \frac{\Delta w}{2} \left(1 - \tanh \left(\frac{z - z_c}{\Delta z} \right) \right) - 1$$

Figure E.1 (Continued)

where:

- z is the cosmological redshift.
- Δw controls the magnitude of the transition in w . The equation of state transitions from $w \approx -1$ at high z to $w \approx -1 + \Delta w$ at low z (specifically, for $z \ll z_c$). Consider Δw in the range $[0, 2]$, ensuring w stays between -1 and 1 .
- z_c is the redshift at which the transition is centered. The prior on z_c is $[0, 0.25]$ and forces the transition to occur in the near past.
- Δz determines the width (in redshift) of the transition. $\log_{10} \Delta z$ is in $[-1.5, 0.5]$.

This parametrization ensures $w(z) \rightarrow -1$ at high redshifts ($z \gg z_c$). The model reduces to a cosmological constant (Λ CDM) if $\Delta w = 0$. It can also appear indistinguishable from Λ CDM if the transition occurs very far in the past ($z_c \gg$ relevant redshifts) or entirely in the future ($z_c \ll 0$ with small Δz).

Parameters / .ini Structure for Added Parameters: This model introduces 3 new parameters beyond the standard Λ CDM parameters. For implementation in CLASS, these parameters should be added to the input `.ini` file. Names and descriptions are:

Figure E.1 (Continued)

Parameter in .ini	Description	Typical Range (Priors)
delta_w	The amplitude of the transition in $w(z)$.	$0 \dots 2$
log10_delta_z	Base-10 logarithm of the width of the transition, $\log_{10}(\Delta z)$.	Fixed to -1.5
zc	The redshift z_c at the center of the transition.	$-2 \dots 2$

Additional background field variables: When running CLASS with this model, the following variables related to the thawing quintessence component should be accessible in the `background.dat` output file:

- w : The equation of state $w(z) = p_{tq}/\rho_{tq}$ of the thawing quintessence fluid.

Implementation Details

- In `background_initial_conditions` you should compute numerically the simple 1d integral $[\ln t_{a_{ini}}^{a_0} 3[(1 + w_{fld})/a]da]$ (e.g. with the Romberg method) instead of calling `background_w_fld` to get `integral_fld` (this is left to 0 in calculations done in `background_w_fld`).
- Use the already implemented minimal Parametrized Post-Friedmann (PPF) approach via the `perturbations.c` module.

HYPOTHESIS GENERATION

In this setting since the LLM has to create its own $(\mathcal{M}, \mathcal{P}_{Fiducial})$, there is an additional auxiliary prompt, `propose_model.md` (Figure E.2). It thus modifies the empty description and $\mathcal{P}_{Fiducial}$ files before proceeding to the remaining stages.

Figure E.2: Prompt entasking the agent to propose a novel cosmological model.

You are a cosmologist, seeking to propose a beyond-Lambda CDM model that will resolve current outstanding issues / tensions in cosmology. Specifically, you want to introduce a model that satisfies the following criteria (in order of importance):

1. Novelty: You should propose a new model that has * not * been implemented / explored in the literature.
2. Your model must optimally satisfy the following four criteria (rewards):
 - a) Minimizes the Hubble Tension (current discrepancies in the H_0 value inferred from late-time model independent observations and "early-universe" measurements)
 - b) Minimizes the S8 tension
 - c) Has a high likelihood with respect to DESI BAO data and a Planck marginal likelihood
 - d) Is consistent with LCDM Cls
3. You will have to implement this model by refactoring the cosmological code CLASS (Cosmic Linear Anisotropy Solving System)** (you are currently in this codebase).

Figure E.2 (Continued)

So you may strategically choose to propose a model that you think would be easier for you to implement in CLASS.

Create a model description as in the example below:

EXAMPLE 1:

EXAMPLE DESCRIPTION.MD: <Example description derived from> *Thawing Quintessence*

EXAMPLE PARAM_BASE.JSON: $\mathcal{P}_{Fiducial}$ dictionary for *Thawing Quintessence*
===

You will need to create a description.md file and a param_base.json file. After doing this, proceed to complete the implementation related tasks in stage1_prompt.md, stage2_prompt.md and stage3_prompt.md. Note, even if the other prompts discourage you from modifying param_base.json, you ****are**** allowed to modify it to to get it to work with your implemented model in CLASS, if for example, it needs extra flags to work with Classy (the Python wrapper to CLASS) successfully.

E.2 ENVIRONMENT DETAILS

The agent performs its tasks in a Podman (rootless) emulator for Docker¹⁴⁰. The Dockerfile that builds the image pre-installs CLASS (version 3.3.0), as well as the code needed to run a Claude Code session. The script that launches the experiment by creating a new container mounts certain host system folders to the containerized environment. This ensures that the agent has access to the nec-

essary auxiliary code and that its submission folder remains on the host system once its experiment ends.

E.2.1 AUXILIARY MODULES AND FILES

The files that are available in the directory, in addition to the base /class code files include prompt files, high-level summaries of key CLASS modules and auxiliary code modules relevant to Stages 2 and 3.

PROMPTS The instructions specific to each stage are separated into three prompt files: `stage1_prompt.md` `stage2_prompt.md`, `stage3_prompt.md`.

FILEMAPS The CLASS code base is considerably long: the `perturbations.c` module, for example, has over 10000 lines of code. To facilitate navigation across such a codebase, we presave LLM-generated summaries of the `background.c`, `perturbations.c` and `input.c` modules. We do so by giving the LLM, Gemini-2.5-pro⁷⁸ a prompt along with the saved module with corresponding line numbers.

E.2.2 AGENT DESIGN

WEB ACCESS The command that launches the agentic experiment disallows the 'WebSearch' and 'WebFetch' tools so the agent cannot use its intrinsic web access tools. The Dockerfile that constructs the environment further contains code that overwrites common web-access relevant commands such as `ping` and `curl` to ensure that they pass through a network filter that only allows "per-

mitted” domains, so as to eliminate the LLM from circumventing the blocked tools by executing bash commands that query the Internet.

CLAUDE CODE We chose Claude Code as the agentic framework since modifications required to implement an arbitrary cosmological model necessitate changes to several parts of the code, to fully incorporate the extension into the code base. Claude Code is a terminal-based agentic framework that has the ability to read files, navigate large code bases and execute code. The npm Claude Code release version we used was 1.0.31. Note, regardless of the version, all experiments we report here used `claude-sonnet-4` as the main model, although the agentic framework relegates easier tasks to a smaller model. By leveraging the ‘allowedTools’ flag while launching a new experiment, we allow most necessary tools, such as bash commands, python commands among others to run without requiring permission from the user. In spite of doing so, the agent will still prompt the human user for permission to run certain risky or new commands (such as a `make` command). This shows up as a multiple-choice selection on the user’s interface. Across experiments we followed the same level of human interaction: we grant it permission for all commands, and do not interact with the agent after the initial experiment-launching prompt.

E.3 PIPELINE DETAILS

An experiment is launched with the following command for the experiments with a provided input cosmological model:

```
claude "Complete the tasks by following the instructions in the three prompts
stage1_prompt.md, stage2_prompt.md and stage3_prompt.md. Start by opening
stage1_prompt.md, completing the tasks described for Stage 1 and then proceed
to complete the tasks for the Stage 2 and Stage 3 by reading the tasks in
those prompts. Your final outputs will be evaluated by the eval_script.sh. You
may NOT access the Internet at any point. Please continue until you have
successfully completed all stages." --model $MODEL --verbose
--disallowedTools="WebSearch,WebFetch"
```

For the *Hypothesis Generation* experiment, the following command launches the agentic experiment:

```
claude "You must first devise your own novel beyond-Lambda CDM cosmological model.
Read the instructions in propose_model_prompt.md. Once you have described your
cosmological model in description.md and param_base.json, complete the tasks
by following the instructions in the three prompts stage1_prompt.md,
stage2_prompt.md and stage3_prompt.md. Start by opening stage1_prompt.md,
completing the tasks described for Stage 1 and then proceed to complete the
tasks for the Stage 2 and Stage 3 by reading the tasks in those prompts. Your
final outputs will be evaluated by the eval_script.sh. You may NOT access the
Internet at any point. Please continue until you have successfully completed
all stages." --model $MODEL --verbose --disallowedTools="WebSearch,WebFetch"
```

E.3.1 STAGE 2

In Stage 2, the LLM must compute observables using $\mathcal{P}_{Fiducial}$. We design the script that the LLM must use to compute observables so that it prioritizes computing the observables via the Python wrapper, `Classy` and subsequently falls back to the C executable if the former fails. The design choice of using a predefined script that the LLM must use to compute observables enforces uniformity in how observables are computed and stored, which enables a straightforward comparison to the target observables.

E.3.2 STAGE 3

In the *Minimization* variant, the agent must create a configuration file that specifies which parameters must be fixed and which should be kept free, and must additionally specify reasonable bounds for each parameter that is to be optimized. In our experiments, we instructed the agent (via the problem context file) to vary the added parameters, in addition to $\{\Omega_{cdm}, h\}$: the density parameter for cold dark matter and the expansion rate of the universe divided by 100. To optimize a different set of parameters, one would only need to modify the single line instruction relevant to this step in the `problem_context.yaml`.

E.4 EXPERIMENT RESULTS

For the experiments we discuss here, the cost of a single experiment with Claude Sonnet-4 is less than \$5: \$3.75 for EDE, \$1.91 for Thawing Quintessence (\$1.92 with the paper as input), and \$4.52

for the Hypothesis Generation experiment with the code that builds successfully.

E.4.1 THAWING QUINTESSENCE

We provide the patch that the LLM generated for the *Thawing Quintessence* experiment below:

Figure E.3: Claude Code's implementation of the *Thawing Quintessence* Model

```
diff --git a/include/background.h b/include/background.h
index fe4d42d8..308bf546 100755
--- a/include/background.h
+++ b/include/background.h
@@ -12,7 +12,7 @@

/** list of possible parametrisations of the DE equation of state */

-enum equation_of_state {CLP,EDE};
+enum equation_of_state {CLP,EDE,TANH};

/** list of possible parametrizations of the varying fundamental constants
    */
@@ -112,6 +112,9 @@ struct background
```

Figure E.3 (Continued)

```
double cs2_fld; /**< \f$ c^2_{s~DE} \f$: sound speed of the fluid in
the frame comoving with the fluid (so, this is
not  $[\delta p/\delta \rho]$  in the synchronous or
newtonian gauge!) */
double Omega_EDE; /**< \f$ w_{DE} \f$: Early Dark Energy density
parameter */
+ double dw_tanh; /**< \f$ \Delta w \f$: amplitude of the
transition in  $w(z)$  for tanh parametrization */
+ double dz_tanh; /**< \f$ \Delta z \f$: width of the transition
for tanh parametrization */
+ double z_tanh; /**< \f$ z_c \f$: redshift at center of
transition for tanh parametrization */
double * scf_parameters; /**< list of parameters describing the scalar
field potential */
short attractor_ic_scf; /**< whether the scalar field has attractor
initial conditions */
int scf_tuning_index; /**< index in scf_parameters used for tuning */
diff --git a/source/background.c b/source/background.c
index f24fc814..ce472491 100755
--- a/source/background.c
+++ b/source/background.c
@@ -702,6 +702,14 @@ int background_w_fld(
```

Figure E.3 (Continued)

```
// w_e(a) taken from eq. (11) in 1706.00730

*w_fld = - d0mega_e(a)/da*a/0mega_e/3./(1.-0mega_e)+a_eq/3./
(a+a_eq);

break;

+ case TANH:
+ {
+ // w(z) = (Delta_w/2) * (1 - tanh((z - z_c)/Delta_z)) - 1
+ // Convert scale factor to redshift: z = 1/a - 1
+ double z = 1./a - 1.;
+ *w_fld = (pba->dw_tanh/2.) * (1. - tanh((z -
+ pba->z_tanh)/pba->dz_tanh)) - 1.;
+ }
+ break;
+ }

@@ -721,6 +729,18 @@ int background_w_fld(
+ d0mega_e(a)/da*d0mega_e(a)/da*a/3./(1.-0mega_e)/
+ (1.-0mega_e)/0mega_e
+ a_eq/3./(a+a_eq)/(a+a_eq);

break;

+ case TANH:
```

Figure E.3 (Continued)

```
+ {
+     // dw/da = dw/dz * dz/da = dw/dz * (-1/a^2)
+     // dw/dz = (Delta_w/2) * (-1/Delta_z) * sech^2((z - z_c)/Delta_z)
+     //          = -(Delta_w/2) * (1/Delta_z) * (1 - tanh^2((z -
+         z_c)/Delta_z))
+     double z = 1./a - 1.;
+     double tanh_arg = (z - pba->z_tanh)/pba->dz_tanh;
+     double tanh_val = tanh(tanh_arg);
+     double dw_dz = -(pba->dw_tanh/2.) * (1. - tanh_val*tanh_val) /
+         pba->dz_tanh;
+     *dw_over_da_fld = dw_dz * (-1./(a*a));
+ }
+ break;
+ }

/** - finally, give the analytic solution of the following integral:
@@ -740,6 +760,10 @@ int background_w_fld(
case EDE:
    class_stop(pba->error_message,"EDE implementation not finished: to
        finish it, read the comments in background.c just before this
        line\n");
    break;
```

Figure E.3 (Continued)

```
+ case TANH:

+ // For TANH, we set integral_fld to 0 and handle it numerically in
    background_initial_conditions

+ *integral_fld = 0.;

+ break;

+ }

    /** note: of course you can generalise these formulas to anything,
@@ -2251,6 +2275,28 @@ int background_initial_conditions(
        [(1+w_fld)/a] da] (e.g. with the Romberg method?) instead of
        calling background_w_fld */

+ /* For TANH parametrization, compute integral numerically if needed */
+ if (pba->fluid_equation_of_state == TANH && integral_fld == 0.) {
+ // Compute integral numerically:  $\int_{a_{ini}}^{a_0} 3(1+w_{fld})/a da$ 
+ // Using simple trapezoidal rule with sufficient steps
+ int n_steps = 1000;
+ double da_step = (1.0 - a) / n_steps;
+ double integral_sum = 0.0;
+ double a_temp, w_temp, dw_temp, integral_temp;
+
+ for (int i = 0; i <= n_steps; i++) {
```

Figure E.3 (Continued)

```
+      a_temp = a + i * da_step;
+      class_call(background_w_fld(pba, a_temp, &w_temp, &dw_temp,
+                                &integral_temp), pba->error_message, pba->error_message);
+      double integrand = 3.0 * (1.0 + w_temp) / a_temp;
+      if (i == 0 || i == n_steps) {
+          integral_sum += 0.5 * integrand * da_step;
+      } else {
+          integral_sum += integrand * da_step;
+      }
+  }
+  integral_fld = integral_sum;
+
+  /* rho_fld at initial time */
+  pvecback_integration[pba->index_bi_rho_fld] = rho_fld_today *
+      exp(integral_fld);
+
+  diff --git a/source/input.c b/source/input.c
+
+  index 933e586c..3e7f30e9 100755
+
+  --- a/source/input.c
+  +++ b/source/input.c
```

Figure E.3 (Continued)

```
@@ -3279,6 +3279,9 @@ int input_read_parameters_species(struct file_content
    * pfc,

        else if ((strstr(string1,"EDE") != NULL) || (strstr(string1,"ede")
            != NULL)) {

            pba->fluid_equation_of_state = EDE;

        }
+    else if ((strstr(string1,"TANH") != NULL) || (strstr(string1,"tanh")
!= NULL)) {
+        pba->fluid_equation_of_state = TANH;
+    }

    else {

        class_stop(errmsg,"incomprehensible input '%s' for the field

            'fluid_equation_of_state'",string1);

    }

@@ -3298,6 +3301,14 @@ int input_read_parameters_species(struct
    file_content * pfc,

        class_read_double("Omega_EDE",pba->Omega_EDE);

        class_read_double("cs2_fld",pba->cs2_fld);

    }

+    if (pba->fluid_equation_of_state == TANH) {
+        /** 8.a.2.4) Equation of state of the fluid in 'TANH' case */
+        /** Read */
```

Figure E.3 (Continued)

```
+     class_read_double("dw_tanh",pba->dw_tanh);
+     class_read_double("dz_tanh",pba->dz_tanh);
+     class_read_double("z_tanh",pba->z_tanh);
+     class_read_double("cs2_fld",pba->cs2_fld);
+ }
}

/** 8.b) If Omega scalar field (SCF) is different from 0 */
@@ -5903,6 +5914,10 @@ int input_default_params(struct background *pba,
    pba->wa_fld = 0.;

    /** 9.a.2.2) 'EDE' case */
    pba->Omega_EDE = 0.;
+ /** 9.a.2.3) 'TANH' case */
+ pba->dw_tanh = 1.0;
+ pba->dz_tanh = 0.03162277660168379; // 10^(-1.5)
+ pba->z_tanh = 0.1;

    /** 9.b) Omega scalar field */

    /** 9.b.1) Potential parameters and initial conditions */
    pba->scf_parameters = NULL;
```

REPEATABILITY

To examine whether the agent is able to consistently solve the *Thawing Quintessence* problem, we ran the experiment 5 times (including the run reported in the main text). We report the median, best and worst results for each of the metrics for the 5 runs below. The ✓ in the first 3 rows indicates that all runs passed those milestones. The agent is thus able to generate code for the cosmological model across all 5 runs. Swapping the style of the input description (as explored in the ‘Paper to Code’ experiment) had a much more significant effect on the results, relative to repeated runs of the same experiment with the curated description.

Table E.1: Results for Repeated *Thawing Quintessence* runs.

Model	Median	Maximum / Worst Result	Minimum / Best Result
Milestone 0 [Patch Exists?]	✓	✓	✓
Milestone 1 [Code compiles?]	✓	✓	✓
Milestone 2 [Observables can be computed?]	✓	✓	✓
$\Omega_{fid}(z)$ Max Frac. Deviation	$4.43e-10$	$2.20e-07$	$9.75e-16$
$mPk(z=0)(k)$ Max Frac. Deviation	$1.31e-4$	$1.37e-04$	$1.15e-04$
$mPk(z=1)(k)$ Max Frac. Deviation	$1.67e-05$	$7.63e-05$	$1.65e-05$
Stage 3 Evaluation	-		
DESI-Planck $\mathcal{P}_{Best-Fit} \ln \mathcal{L}$	-14.7	-14.7	-14.7

References

- [1] (n.d.). Jpype. Available online at: <https://jpype.readthedocs.io/en/latest/>.
- [2] Abbott, T. M., Agüena, M., Alarcon, A., Allam, S., Alves, O., Amon, A., Andrade-Oliveira, F., Annis, J., Avila, S., Bacon, D., et al. (2022). Dark energy survey year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105(2), 023520.
- [3] Abergel, A., Ade, P. A., Aghanim, N., Alves, M., Aniano, G., Armitage-Caplan, C., Arnaud, M., Ashdown, M., Atrio-Barandela, F., Aumont, J., et al. (2014). Planck 2013 results. xi. all-sky model of thermal dust emission. *Astronomy & Astrophysics*, 571, A11.
- [4] Abolfathi, B., Aguado, D., Aguilar, G., Prieto, C. A., Almeida, A., Ananna, T. T., Anders, F., Anderson, S. F., Andrews, B. H., Anguiano, B., et al. (2018). The fourteenth data release of the sloan digital sky survey: First spectroscopic data from the extended baryon oscillation spectroscopic survey and from the second phase of the apache point observatory galactic evolution experiment. *The Astrophysical Journal Supplement Series*, 235(2), 42.
- [5] Accetta, K., Aerts, C., Aguirre, V. S., Ahumada, R., Ajgaonkar, N., Ak, N. F., Alam, S., Prieto, C. A., Almeida, A., Anders, F., et al. (2022). The seventeenth data release of the sloan digital sky surveys: Complete release of manga, mastar, and apogee-2 data. *The Astrophysical Journal Supplement Series*, 259(2), 35.
- [6] Adam, R., Ade, P., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A., Barreiro, R., Bartolo, N., et al. (2016). Planck 2015 results-viii. high frequency instrument data processing: Calibration and maps. *Astronomy & Astrophysics*, 594, A8.
- [7] Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Anderson, K. S., Anderson, S. F., Annis, J., Bahcall, N. A., Bailer-Jones, C. A., Baldry, I. K., Barentine, J. C., et al. (2007). The fifth data release of the sloan digital sky survey. *The Astrophysical Journal Supplement Series*, 172(2), 634.
- [8] Aghanim, N., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A., Barreiro, R., Bartolo, N., Basak, S., Benabed, K., et al. (2016). Planck intermediate results-xlviii.

- disentangling galactic dust emission and cosmic infrared background anisotropies. *Astronomy & Astrophysics*, 596, A109.
- [9] Akeson, R., Armus, L., Bachelet, E., Bailey, V., Bartusek, L., Bellini, A., Benford, D., Bennett, D., Bhattacharya, A., Bohlin, R., et al. (2019). The wide field infrared survey telescope: 100 hubbles for the 2020s. *arXiv preprint arXiv:1902.05569*.
 - [10] Anderson, B. D. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3), 313–326.
 - [11] Andrae, R., Fouesneau, M., Sordo, R., Bailer-Jones, C., Dharmawardena, T., Rybizki, J., De Angeli, F., Lindstrøm, H., Marshall, D., Drimmel, R., et al. (2022). Gaia data release 3: Analysis of the gaia bp/rp spectra using the general stellar parameterizer from photometry. *arXiv preprint arXiv:2206.06138*.
 - [12] Ansys, Inc. (n.d.). Ansys. Available online at: <https://www.ansys.com/>.
 - [13] Anthropic. Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>.
 - [14] Anthropic. Claude code. <https://docs.anthropic.com/en/docs/claude-code/overview>.
 - [15] Anthropic. Claude sonnet 4. <https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf>.
 - [16] Arce, H. G. & Goodman, A. A. (1999). Measuring galactic extinction: A test. *The Astrophysical Journal*, 512(2), L135.
 - [17] Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., Lim, P. L., Crawford, S. M., Conseil, S., Shupe, D. L., Craig, M. W., Dencheva, N., Ginsburg, A., VanderPlas, J. T., Bradley, L. D., Pérez-Suárez, D., de Val-Borro, M., Aldcroft, T. L., Cruz, K. L., Robitaille, T. P., Tollerud, E. J., Ardelean, C., Babej, T., Bach, Y. P., Bachetti, M., Bakanov, A. V., Bamford, S. P., Barentsen, G., Barmby, P., Baumbach, A., Berry, K. L., Biscani, F., Boquien, M., Bostroem, K. A., Bouma, L. G., Brammer, G. B., Bray, E. M., Breytenbach, H., Buddelmeijer, H., Burke, D. J., Calderone, G., Cano Rodríguez, J. L., Cara, M., Cardoso, J. V. M., Cheedella, S., Copin, Y., Corrales, L., Crichton, D., D’Avella, D., Deil, C., Depagne, É., Dietrich, J. P., Donath, A., Droettboom, M., Earl, N., Erben, T., Fabbro, S., Ferreira,

- L. A., Finethy, T., Fox, R. T., Garrison, L. H., Gibbons, S. L. J., Goldstein, D. A., Gommers, R., Greco, J. P., Greenfield, P., Groener, A. M., Grollier, F., Hagen, A., Hirst, P., Homeier, D., Horton, A. J., Hosseinzadeh, G., Hu, L., Hunkeler, J. S., Ivezić, Ž., Jain, A., Jenness, T., Kanarek, G., Kendrew, S., Kern, N. S., Kerzendorf, W. E., Khvalko, A., King, J., Kirkby, D., Kulkarni, A. M., Kumar, A., Lee, A., Lenz, D., Littlefair, S. P., Ma, Z., Macleod, D. M., Mastropietro, M., McCully, C., Montagnac, S., Morris, B. M., Mueller, M., Mumford, S. J., Muna, D., Murphy, N. A., Nelson, S., Nguyen, G. H., Ninan, J. P., Nöthe, M., Ogaz, S., Oh, S., Parejko, J. K., Parley, N., Pascual, S., Patil, R., Patil, A. A., Plunkett, A. L., Prochaska, J. X., Rastogi, T., Reddy Janga, V., Sabater, J., Sakurikar, P., Seifert, M., Sherbert, L. E., Sherwood-Taylor, H., Shih, A. Y., Sick, J., Silbiger, M. T., Singanamalla, S., Singer, L. P., Sladen, P. H., Sooley, K. A., Sornarajah, S., Streicher, O., Teuben, P., Thomas, S. W., Tremblay, G. R., Turner, J. E. H., Terrón, V., van Kerkwijk, M. H., de la Vega, A., Watkins, L. L., Weaver, B. A., Whitmore, J. B., Woillez, J., Zabalza, V., & Astropy Contributors (2018). The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *The Astronomical Journal*, 156(3), 123.
- [18] Audenaert, J., Muthukrishna, D., Gregory, P. F., Hogg, D. W., & Villar, V. A. (2025). Causal foundation models: Disentangling physics from instrument properties. In *1st ICML Workshop on Foundation Models for Structured Data*.
- [19] Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. (2021). Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.
- [20] Awan, H., Gawiser, E., Kurczynski, P., Jones, R. L., Zhan, H., Padilla, N. D., Arancibia, A. M. M., Orsi, A., Cora, S. A., & Yoachim, P. (2016). Testing lsst dither strategies for survey uniformity and large-scale structure systematics. *The Astrophysical Journal*, 829(1), 50.
- [21] Bansal, A., Borgnia, E., Chu, H.-M., Li, J., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., & Goldstein, T. (2024). Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36.
- [22] Baratta, I. A., Dean, J. P., Dokken, J. S., Habera, M., Hale, J. S., Richardson, C. N., Rognes, M. E., Scroggs, M. W., Sime, N., & Wells, G. N. (2023). DOLFINx: the next generation FEniCS problem solving environment. *preprint*.

- [23] Basu, K., Abdelaziz, I., Bradford, K., Crouse, M., Kate, K., Kumaravel, S., Goyal, S., Munawar, A., Rizk, Y., Wang, X., et al. (2024). Nestful: A benchmark for evaluating llms on nested sequences of api calls. *arXiv preprint arXiv:2409.03797*.
- [24] Beichman, C., Neugebauer, G., Habing, H., Clegg, P., & Chester, T. J. (1988). *Infrared astronomical satellite (IRAS) catalogs and atlases. Volume 1: Explanatory supplement*. Technical report.
- [25] Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.
- [26] Biewald, L. (2020). Experiment tracking with weights and biases. Software available from wandb.com.
- [27] Blanton, M. R., Bershad, M. A., Abolfathi, B., Albareti, F. D., Prieto, C. A., Almeida, A., Alonso-García, J., Anders, F., Anderson, S. F., Andrews, B., et al. (2017). Sloan digital sky survey iv: Mapping the milky way, nearby galaxies, and the distant universe. *The Astronomical Journal*, 154(1), 28.
- [28] Blanton, M. R., Schlegel, D. J., Strauss, M. A., Brinkmann, J., Finkbeiner, D., Fukugita, M., Gunn, J. E., Hogg, D. W., Ivezić, Ž., Knapp, G., et al. (2005). New york university value-added galaxy catalog: A galaxy catalog based on new public surveys. *The Astronomical Journal*, 129(6), 2562.
- [29] Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., et al. (2025). A foundation model for the earth system. *Nature*, 641(8065), 1180–1187.
- [30] Boelens, A. M. & Tchelepi, H. A. (2021). Quantimpy: Minkowski functionals and functions with python. *SoftwareX*, 16, 100823.
- [31] Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiušė, K., Askill, A., Jones, A., Chen, A., et al. (2022). Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.
- [32] Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2023). Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

- [33] Bravo, M., Gawiser, E., Padilla, N. D., DeRose, J., Wechsler, R. H., Collaboration, L. D. E. S., et al. (2021). Simultaneous estimation of large-scale structure and milky way dust extinction from galaxy surveys. *The Astrophysical Journal*, 921(2), 108.
- [34] Brown, A., Vallenari, A., Prusti, T., De Bruijne, J., Babusiaux, C., Bailer-Jones, C., Biermann, M., Evans, D. W., Eyer, L., Jansen, F., et al. (2018). Gaia data release 2-summary of the contents and survey properties. *Astronomy & Astrophysics*, 616, A1.
- [35] Brown, A. G., Vallenari, A., Prusti, T., De Bruijne, J., Babusiaux, C., Biermann, M., Creevey, O., Evans, D., Eyer, L., Hutton, A., et al. (2021). Gaia early data release 3-summary of the contents and survey properties. *Astronomy & Astrophysics*, 649, A1.
- [36] Burstein, D. & Heiles, C. (1978). Hi, galaxy counts, and reddening-variation in the gas-to-dust ratio, the extinction at high galactic latitudes, and a new method for determining galactic reddening. *The Astrophysical Journal*, 225, 40–55.
- [37] Cassano, F., Gouwar, J., Lucchetti, F., Schlesinger, C., Freeman, A., Anderson, C. J., Feldman, M. Q., Greenberg, M., Jangda, A., & Guha, A. (2024). Knowledge transfer from high-resource to low-resource programming languages for code llms. *Proceedings of the ACM on Programming Languages*, 8(OOPSLA2), 677–708.
- [38] Chambers, K. C., Magnier, E., Metcalfe, N., Flewelling, H., Huber, M., Waters, C., Denneau, L., Draper, P., Farrow, D., Finkbeiner, D., et al. (2016). The pan-starrs1 surveys. *arXiv preprint arXiv:1612.05560*.
- [39] Chase, H. (2022). LangChain.
- [40] Chen, H., Dong, Y., Shao, S., Hao, Z., Yang, X., Su, H., & Zhu, J. (2024). Your diffusion model is secretly a certifiably robust classifier. *arXiv preprint arXiv:2402.02316*.
- [41] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- [42] Chen, S.-F., White, M., DeRose, J., & Kokron, N. (2022). Cosmological analysis of three-dimensional boss galaxy clustering and planck cmb lensing cross correlations via lagrangian perturbation theory. *arXiv preprint arXiv:2204.10392*.
- [43] Chiang, T. (2000). Catching crumbs from the table. *Nature*, 405(6786), 517–517.

- [44] Chiang, Y.-K. (2023). Corrected sfid: A more accurate galactic dust map with minimal extragalactic contamination. *The Astrophysical Journal*, 958(2), 118.
- [45] Chiang, Y.-K. & Ménard, B. (2019). Extragalactic imprints in galactic dust maps. *The Astrophysical Journal*, 870(2), 120.
- [46] Chung, D. J., Gao, Z., Kvasiuk, Y., Li, T., Münchmeyer, M., Rudolph, M., Sala, F., & Tadepalli, S. C. (2025). Theoretical physics benchmark (tpbench)—a dataset and study of ai reasoning capabilities in theoretical physics. *arXiv preprint arXiv:2502.15815*.
- [47] Clark, K. & Jaini, P. (2024). Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36.
- [48] Cobb, A. D., Baydin, A. G., Markham, A., & Roberts, S. J. (2019). Introducing an explicit symplectic integration scheme for riemannian manifold hamiltonian monte carlo. *arXiv preprint arXiv:1910.06243*.
- [49] CodeGemma Team, Zhao, H., Hui, J., Howland, J., Nguyen, N., Zuo, S., Hu, A., Choquette-Choo, C. A., Shen, J., Kelley, J., et al. (2024). Codegemma: Open code models based on gemma. *arXiv preprint arXiv:2406.11409*.
- [50] Collaboration, D. E. S. I., Aghamousa, A., Aguilar, J., Ahlen, S., Alam, S., Allen, L. E., Prieto, C. A., Annis, J., Bailey, S., Balland, C., et al. (2016a). The desi experiment part i: Science, targeting, and survey design.
- [51] Collaboration, P. et al. (2016b). Planck 2015 results: Vi. lfi mapmaking. *Astronomy & Astrophysics*, (pp. A6).
- [52] COMSOL Multiphysics®(n.d.). COMSOL Multiphysics®. Available online at: <https://www.comsol.com/>.
- [53] COMSOL Multiphysics® (n.d.). COMSOL Multiphysics® application gallery. Available online at: <https://www.comsol.com/models>.
- [54] Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. (2022). Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*.
- [55] Courant, R. et al. (1994). Variational methods for the solution of problems of equilibrium and vibrations. *Lecture notes in pure and applied mathematics*, (pp. 1–1).

- [56] Cuesta-Lazaro, C. & Mishra-Sharma, S. (2023). A point cloud approach to generative modeling for galaxy surveys at the field level. *arXiv preprint arXiv:2311.17141*.
- [57] Cutri, R., Skrutskie, M., Van Dyk, S., Beichman, C., Carpenter, J., Chester, T., Cambresy, L., Evans, T., et al. (2003a). Explanatory supplement to the 2mass all sky data release.
- [58] Cutri, R., Skrutskie, M., Van Dyk, S., Beichman, C., Carpenter, J., Chester, T., Cambresy, L., Evans, T., Fowler, J., Gizis, J., et al. (2003b). 2mass all sky catalog of point sources. *The IRSA 2MASS All-Sky Point Source Catalog*.
- [59] Cutri, R., Skrutskie, M., Van Dyk, S., Beichman, C., Carpenter, J., Chester, T., Cambresy, L., Evans, T., Fowler, J., Gizis, J., et al. (2006). Explanatory supplement to the 2mass all sky data release and extended mission products. *Caltech, Pasadena*.
- [60] Cutri, R., Wright, E., Conrow, T., Fowler, J., Eisenhardt, P., Grillmair, C., Kirkpatrick, J., Masci, F., McCallon, H., Wheelock, S., et al. (2013). Explanatory supplement to the allwise data release products. *Explanatory Supplement to the AllWISE Data Release Products*, (pp.1).
- [61] Dai, B. & Seljak, U. (2024). Multiscale flow for robust and optimal cosmological analysis. *Proceedings of the National Academy of Sciences*, 121(9), e2309624121.
- [62] Dassault Systèmes (n.d.). Abaqus, fea. Available online at: <https://www.3ds.com/products/simulia/abaqus>.
- [63] Delchambre, L., Bailer-Jones, C., Bellas-Velidis, I., Drimmel, R., Garabato, D., Carballo, R., Hatzidimitriou, D., Marshall, D., Andrae, R., Dafonte, C., et al. (2022). Gaia dr3: Apsis iii–non-stellar content and source classification. *arXiv preprint arXiv:2206.06710*.
- [64] DESI Collaboration, Abdul-Karim, M., Aguilar, J., Ahlen, S., Alam, S., Allen, L., Prieto, C. A., Alves, O., Anand, A., Andrade, U., Armengaud, E., Aviles, A., Bailey, S., Baltay, C., Bansal, P., Bault, A., Behera, J., BenZvi, S., Bianchi, D., Blake, C., Brieden, S., Brodzeller, A., Brooks, D., Buckley-Geer, E., Burtin, E., Calderon, R., Canning, R., Rosell, A. C., Carrilho, P., Casas, L., Castander, F. J., Cereskaite, R., Charles, M., Chaussidon, E., Chaves-Montero, J., Chebat, D., Chen, X., Claybaugh, T., Cole, S., Cooper, A. P., Cuceu, A., Dawson, K. S., de la Macorra, A., de Mattia, A., Deiosso, N., Costa, J. D., Demina, R., Dey, A., Dey, B., Ding, Z., Doel, P., Edelstein, J., Eisenstein, D. J., Elbers, W., Fagrelus, P., Fanning, K., Fernández-García, E., Ferraro, S., Font-Ribera, A., Forero-Romero, J. E., Frenk, C. S., Garcia-Quintero, C., Garrison, L. H., Gaztañaga, E., Gil-Marín, H., Gontcho, S. G. A., Gonzalez,

D., Gonzalez-Morales, A. X., Gordon, C., Green, D., Gutierrez, G., Guy, J., Hadzhiyska, B., Hahn, C., He, S., Herbold, M., Herrera-Alcantar, H. K., Ho, M., Honscheid, K., Howlett, C., Huterer, D., Ishak, M., Juneau, S., Kamble, N. V., Karaçaylı, N. G., Kehoe, R., Kent, S., Kim, A. G., Kirkby, D., Kisner, T., Koposov, S. E., Kremin, A., Krolewski, A., Lahav, O., Lamman, C., Landriau, M., Lang, D., Lasker, J., Goff, J. M. L., Guillou, L. L., Leauthaud, A., Levi, M. E., Li, Q., Li, T. S., Lodha, K., Lokken, M., Lozano-Rodríguez, F., Magneville, C., Manera, M., Martini, P., Matthewson, W. L., Meisner, A., Mena-Fernández, J., Menegas, A., Mergulhão, T., Miquel, R., Moustakas, J., Muñoz-Gutiérrez, A., Muñoz-Santos, D., Myers, A. D., Nadathur, S., Naidoo, K., Napolitano, L., Newman, J. A., Niz, G., Noriega, H. E., Paillas, E., Palanque-Delabrouille, N., Pan, J., Peacock, J., Ibanez, M. P., Percival, W. J., Pérez-Fernández, A., Pérez-Ràfols, I., Pieri, M. M., Poppett, C., Prada, F., Rabinowitz, D., Raichoor, A., Ramírez-Pérez, C., Rashkovetskyi, M., Ravoux, C., Rich, J., Rocher, A., Rockosi, C., Rohlf, J., Román-Herrera, J. O., Ross, A. J., Rossi, G., Ruggeri, R., Ruhlmann-Kleider, V., Samushia, L., Sanchez, E., Sanders, N., Schlegel, D., Schubnell, M., Seo, H., Shafieloo, A., Sharples, R., Silber, J., Sinigaglia, F., Sprayberry, D., Tan, T., Tarlé, G., Taylor, P., Turner, W., Ureña-López, L. A., Vaisakh, R., Valdes, F., Valogiannis, G., Vargas-Magaña, M., Verde, L., Walther, M., Weaver, B. A., Weinberg, D. H., White, M., Wolfson, M., Yèche, C., Yu, J., Zaborowski, E. A., Zarrouk, P., Zhai, Z., Zhang, H., Zhao, C., Zhao, G. B., Zhou, R., & Zou, H. (2025). Desi dr2 results ii: Measurements of baryon acoustic oscillations and cosmological constraints.

- [65] Dhariwal, P. & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 8780–8794.
- [66] Di Valentino, E., Mena, O., Pan, S., Visinelli, L., Yang, W., Melchiorri, A., Mota, D. F., Riess, A. G., & Silk, J. (2021). In the realm of the hubble tension—a review of solutions. *Classical and Quantum Gravity*, 38(15), 153001.
- [67] Dodelson, S., Heitmann, K., Hirata, C., Honscheid, K., Roodman, A., Seljak, U., Slosar, A., & Trodden, M. (2016). Cosmic visions dark energy: Science. *arXiv preprint arXiv:1604.07626*.
- [68] Dole, H., Lagache, G., Puget, J.-L., Caputi, K. I., Fernandez-Conde, N., Le Floc’h, E., Papovich, C., Pérez-González, P. G., Rieke, G. H., & Blaylock, M. (2006). The cosmic infrared background resolved by spitzer-contributions of mid-infrared galaxies to the far-infrared background. *Astronomy & Astrophysics*, 451(2), 417–429.

- [69] Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, 195(2), 216–222.
- [70] Edenhofer, G., Zucker, C., Frank, P., Saydjari, A. K., Speagle, J. S., Finkbeiner, D., & Enßlin, T. A. (2024). A parsec-scale galactic 3d dust map out to 1.25 kpc from the sun. *Astronomy & Astrophysics*, 685, A82.
- [71] Feder, R. M., Berger, P., & Stein, G. (2020). Nonlinear 3d cosmic web simulation with heavy-tailed generative adversarial networks. *Phys. Rev. D*, 102, 103504.
- [72] Finkbeiner, D. P., Schlafly, E. F., Schlegel, D. J., Padmanabhan, N., Jurić, M., Burgett, W. S., Chambers, K. C., Denneau, L., Draper, P. W., Flewelling, H., et al. (2016). Hypercalibration: a pan-starrs1-based recalibration of the sloan digital sky survey photometry. *The Astrophysical Journal*, 822(2), 66.
- [73] Flewelling, H., Magnier, E., Chambers, K., Heasley, J., Holmberg, C., Huber, M., Sweeney, W., Waters, C., Calamida, A., Casertano, S., et al. (2020). The pan-starrs1 database and data products. *The Astrophysical Journal Supplement Series*, 251(1), 7.
- [74] GaiaCollaboration (2016). The gaia mission. *arXiv preprint arXiv:1609.04153*.
- [75] Garcia-Fernandez, M., Sanchez, E., Sevilla-Noarbe, I., Suchyta, E., Huff, E., Gaztanaga, E., Aleksić, J., Ponce, R., Castander, F., Hoyle, B., et al. (2018). Weak lensing magnification in the dark energy survey science verification data. *Monthly Notices of the Royal Astronomical Society*, 476(1), 1071–1085.
- [76] Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. (2024). Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- [77] Google Cloud Vertex AI (n.d.). Discoveryengine.
- [78] Google LLC (n.d.). Gemini 2.5 pro model card. <https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf>.
- [79] Gorski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., & Bartelmann, M. (2005). Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2), 759.

- [80] Green, G. (2018). dustmaps: A Python interface for maps of interstellar dust. *The Journal of Open Source Software*, 3(26), 695.
- [81] Green, G. M., Schlafly, E., Zucker, C., Speagle, J. S., & Finkbeiner, D. (2019a). A 3d dust map based on gaia, pan-starrs 1, and 2mass. *The Astrophysical Journal*, 887(1), 93.
- [82] Green, G. M., Schlafly, E., Zucker, C., Speagle, J. S., & Finkbeiner, D. (2019b). A 3d dust map based on gaia, pan-starrs 1, and 2mass. *The Astrophysical Journal*, 887(1), 93.
- [83] Green, G. M., Schlafly, E. F., Finkbeiner, D., Rix, H.-W., Martin, N., Burgett, W., Draper, P. W., Flewelling, H., Hodapp, K., Kaiser, N., et al. (2018). Galactic reddening in 3d from stellar photometry—an improved map. *Monthly Notices of the Royal Astronomical Society*, 478(1), 651–666.
- [84] Green, G. M., Schlafly, E. F., Finkbeiner, D. P., Jurić, M., Rix, H.-W., Burgett, W., Chambers, K. C., Draper, P. W., Flewelling, H., Kudritzki, R. P., et al. (2014). Measuring distances and reddenings for a billion stars: Toward a 3d dust map from pan-starrs 1. *The Astrophysical Journal*, 783(2), 114.
- [85] Green, G. M., Schlafly, E. F., Finkbeiner, D. P., Rix, H. W., Martin, N., Burgett, W., Draper, P. W., Flewelling, H., Hodapp, K., Kaiser, N., Kudritzki, R. P., Magnier, E., Metcalfe, N., Price, P., Tonry, J., & Wainscoat, R. (2015). a three-dimensional map of milky way dust. *Astrophysical Journal*, 810.
- [86] Gunn, J. E., Siegmund, W. A., Mannery, E. J., Owen, R. E., Hull, C. L., Leger, R. F., Carey, L. N., Knapp, G. R., York, D. G., Boroski, W. N., et al. (2006). The 2.5 m telescope of the sloan digital sky survey. *The Astronomical Journal*, 131(4), 2332.
- [87] Hahn, C., Eickenberg, M., Ho, S., Hou, J., Lemos, P., Massara, E., Modi, C., Moradinezhad Dizgah, A., Blancard, B. R.-S., & Abidi, M. M. (2023). A forward modeling approach to analyzing galaxy clustering with simbig. *Proceedings of the National Academy of Sciences*, 120(42), e2218810120.
- [88] Hamaus, N., Pisani, A., Sutter, P. M., Lavaux, G., Escoffier, S., Wandelt, B. D., & Weller, J. (2016). Constraints on cosmology and gravity from the dynamics of voids. *Physical Review Letters*, 117(9), 091302.
- [89] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk,

- M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362.
- [90] Hartlap, J., Simon, P., & Schneider, P. (2007). Why your model parameter confidences might be too optimistic. unbiased estimation of the inverse covariance matrix. *Astronomy & Astrophysics*, 464(1), 399–404.
- [91] Hauser, M., Kelsall, T., Leisawitz, D., & Weiland, J. (1998). COBE diffuse infrared background experiment (DIRBE) explanatory supplement. *Greenbelt: NASA*.
- [92] Heitmann, K., Higdon, D., White, M., Habib, S., Williams, B. J., & Wagner, C. (2009). The Coyote Universe II: Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum. *Astrophys. J.*, 705, 156–174.
- [93] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- [94] Hennig, J., Elfner, M., Andreas, M., & Jacob, F. (n.d.). Mph. Available online at: <https://mph.readthedocs.io/en/1.2/>.
- [95] Heurtel-Depeiges, D., Burkhart, B., Ohana, R., & Blancard, B. R.-S. (2023). Removing dust from cmb observations with diffusion models. *arXiv preprint arXiv:2310.16285*.
- [96] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- [97] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
- [98] Horowitz, B. & Melchior, P. (2022). Plausible adversarial attacks on direct parameter inference models in astrophysics. *arXiv preprint arXiv:2211.14788*.
- [99] Huang, Z., Johnson, J., Debnath, S., Rehg, J. M., & Wu, C.-Y. (2024). Pointinfinity: Resolution-invariant point diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10050–10060).

- [100] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3), 90–95.
- [101] Indebetouw, R., Mathis, J., Babler, B., Meade, M., Watson, C., Whitney, B., Wolff, M., Wolfire, M., Cohen, M., Bania, T., et al. (2005). The wavelength dependence of interstellar extinction from 1.25 to 8.0 μm using glimpse data. *The Astrophysical Journal*, 619(2), 931.
- [102] Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., et al. (2019). Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2), 111.
- [103] Ivezić, Ž., Sesar, B., Jurić, M., Bond, N., Dalcanton, J., Rockosi, C. M., Yanny, B., Newberg, H. J., Beers, T. C., Prieto, C. A., et al. (2008). The milky way tomography with sdss. ii. stellar metallicity. *The Astrophysical Journal*, 684(1), 287.
- [104] Jagvaral, Y., Mandelbaum, R., & Lanusse, F. (2022). Modeling halo and central galaxy orientations on the so (3) manifold with score-based generative models. *arXiv preprint arXiv:2212.05592*.
- [105] Jamieson, D., Li, Y., de Oliveira, R. A., Villaescusa-Navarro, F., Ho, S., & Spergel, D. N. (2022). Field level neural network emulator for cosmological n-body simulations. *arXiv preprint arXiv:2206.04594*.
- [106] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2023). Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.
- [107] Juric, M. (2012). Lsd: Large survey database framework. *Astrophysics Source Code Library*, (pp. ascl-1209).
- [108] Kamionkowski, M. & Riess, A. G. (2023). The Hubble Tension and Early Dark Energy. *Ann. Rev. Nucl. Part. Sci.*, 73, 153–180.
- [109] Karim, T., Singh, S., Rezaie, M., Eisenstein, D., Hadzhiyska, B., Speagle, J. S., Aguilar, J. N., Ahlen, S., Brooks, D., Claybaugh, T., et al. (2025). Measuring σ_8 using desi legacy imaging surveys emission-line galaxies and planck cmb lensing, and the impact of dust on parameter inference. *Journal of Cosmology and Astroparticle Physics*, 2025(02), 045.

- [110] Karwal, T. & Kamionkowski, M. (2016). Dark energy at early times, the Hubble parameter, and the string axiverse. *Phys. Rev. D*, 94(10), 103523.
- [111] Kasim, M., Watson-Parris, D., Deaconu, L., Oliver, S., Hatfield, P., Froula, D., Gregori, G., Jarvis, M., Khatiwala, S., Korenaga, J., et al. (2021). Building high accuracy emulators for scientific simulations with deep neural architecture search. *Machine Learning: Science and Technology*, 3(1), 015013.
- [112] Kingma, D., Salimans, T., Poole, B., & Ho, J. (2021). Variational diffusion models. *Advances in neural information processing systems*, 34, 21696–21707.
- [113] Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [114] Kingma, D. P. & Gao, R. (2023). Understanding the diffusion objective as a weighted integral of elbos. *arXiv preprint arXiv:2303.00848*.
- [115] Kitanidis, E., White, M., Feng, Y., Schlegel, D., Guy, J., Dey, A., Landriau, M., Brooks, D., Levi, M., Moustakas, J., et al. (2020). Imaging systematics and clustering of desi main targets. *Monthly Notices of the Royal Astronomical Society*, 496(2), 2262–2291.
- [116] Krause, E., Eifler, T., Zuntz, J., Friedrich, O., Troxel, M., Dodelson, S., Blazek, J., Secco, L., MacCrann, N., Baxter, E., et al. (2017). Dark energy survey year 1 results: multi-probe methodology and simulated likelihood analyses. *arXiv preprint arXiv:1706.09359*.
- [117] Kulal, S., Pasupat, P., Chandra, K., Lee, M., Padon, O., Aiken, A., & Liang, P. S. (2019). Spoc: Search-based pseudocode to code. *Advances in Neural Information Processing Systems*, 32.
- [118] Kumar, V., Gleyzer, L., Kahana, A., Shukla, K., & Karniadakis, G. E. (2023). My-crunchgpt: A chatgpt assisted framework for scientific machine learning. *arXiv preprint arXiv:2306.15551*.
- [119] Kumarappan, A., Tiwari, M., Song, P., George, R. J., Xiao, C., & Anandkumar, A. (2024). Leanagent: Lifelong learning for formal theorem proving. *arXiv preprint arXiv:2410.06209*.
- [120] Lagache, G., Puget, J.-L., & Dole, H. (2005). Dusty infrared galaxies: sources of the cosmic infrared background. *Annu. Rev. Astron. Astrophys.*, 43, 727–768.

- [121] Lallement, R., Babusiaux, C., Vergely, J., Katz, D., Arenou, F., Valette, B., Hottier, C., & Capitanio, L. (2019). Gaia-2mass 3d maps of galactic interstellar dust within 3 kpc. *Astronomy & Astrophysics*, 625, A135.
- [122] Legin, R., Ho, M., Lemos, P., Perreault-Levasseur, L., Ho, S., Hezaveh, Y., & Wandelt, B. (2024). Posterior sampling of the initial conditions of the universe from non-linear large scale structures using score-based generative models. *Monthly Notices of the Royal Astronomical Society: Letters*, 527(1), L173–L178.
- [123] Leike, R. & Enßlin, T. (2019). Charting nearby dust clouds using gaia data only. *Astronomy & Astrophysics*, 631, A32.
- [124] Leike, R., Glatzle, M., & Enßlin, T. (2020). Resolving nearby dust clouds. *Astronomy & Astrophysics*, 639, A138.
- [125] Lenz, D., Hensley, B. S., & Doré, O. (2017). A new, large-scale map of interstellar reddening derived from h i emission. *The Astrophysical Journal*, 846(1), 38.
- [126] Lesgourgues, J. (2011). The cosmic linear anisotropy solving system (class) i: Overview.
- [127] Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. (2022). Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35, 3843–3857.
- [128] Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., & Pathak, D. (2023). Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2206–2217).
- [129] Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al. (2022). Competition-level code generation with alphacode. *Science*, 378(6624), 1092–1097.
- [130] Lodha, K., Calderon, R., Matthewson, W. L., Shafieloo, A., Ishak, M., Pan, J., Garcia-Quintero, C., Huterer, D., Valogiannis, G., Ureña-López, L. A., Kamble, N. V., Parkinson, D., Kim, A. G., Zhao, G. B., Cervantes-Cota, J. L., Rohlf, J., Lozano-Rodríguez, F., Román-Herrera, J. O., Abdul-Karim, M., Aguilar, J., Ahlen, S., Alves, O., Andrade, U., Armengaud, E., Aviles, A., BenZvi, S., Bianchi, D., Brodzeller, A., Brooks, D., Burtin, E., Canning, R.,

Rosell, A. C., Casas, L., Castander, F. J., Charles, M., Chaussidon, E., Chaves-Montero, J., Chebat, D., Claybaugh, T., Cole, S., Cuceu, A., Dawson, K. S., de la Macorra, A., de Mattia, A., Deiosso, N., Demina, R., Dey, A., Dey, B., Ding, Z., Doel, P., Eisenstein, D. J., Elbers, W., Ferraro, S., Font-Ribera, A., Forero-Romero, J. E., Garrison, L. H., Gaztañaga, E., Gil-Marín, H., Gontcho, S. G. A., Gonzalez-Morales, A. X., Gutierrez, G., Guy, J., Hahn, C., Herbold, M., Herrera-Alcantar, H. K., Honscheid, K., Howlett, C., Juneau, S., Kehoe, R., Kirkby, D., Kisner, T., Kremin, A., Lahav, O., Lamman, C., Landriau, M., Guillo, L. L., Leauthaud, A., Levi, M. E., Li, Q., Magneville, C., Manera, M., Martini, P., Meisner, A., Mena-Fernández, J., Miquel, R., Moustakas, J., Santos, D. M., Muñoz-Gutiérrez, A., Myers, A. D., Nadathur, S., Niz, G., Noriega, H. E., Paillas, E., Palanque-Delabrouille, N., Percival, W. J., Pieri, M. M., Poppett, C., Prada, F., Pérez-Fernández, A., Pérez-Ràfols, I., Ramírez-Pérez, C., Rashkovetskyi, M., Ravoux, C., Ross, A. J., Rossi, G., Ruhlmann-Kleider, V., Samushia, L., Sanchez, E., Schlegel, D., Schubnell, M., Seo, H., Sinigaglia, F., Sprayberry, D., Tan, T., Tarlé, G., Taylor, P., Turner, W., Vargas-Magaña, M., Walther, M., Weaver, B. A., Wolfson, M., Yèche, C., Zarrouk, P., Zhou, R., & Zou, H. (2025). Extended dark energy analysis using desi dr2 bao measurements.

- [131] Magnier, E. A., Chambers, K., Flewelling, H., Hoblitt, J., Huber, M., Price, P., Sweeney, W., Waters, C., Denneau, L., Draper, P., et al. (2020a). The pan-starrs data-processing system. *The Astrophysical Journal Supplement Series*, 251(1), 3.
- [132] Magnier, E. A., Schlafly, E. F., Finkbeiner, D. P., Tonry, J., Goldman, B., Röser, S., Schilbach, E., Casertano, S., Chambers, K., Flewelling, H., et al. (2020b). Pan-starrs photometric and astrometric calibration. *The Astrophysical Journal Supplement Series*, 251(1), 6.
- [133] Magnier, E. A., Sweeney, W., Chambers, K., Flewelling, H., Huber, M., Price, P., Waters, C., Denneau, L., Draper, P., Farrow, D., et al. (2020c). Pan-starrs pixel analysis: source detection and characterization. *The Astrophysical Journal Supplement Series*, 251(1), 5.
- [134] Mainzer, A., Bauer, J., Grav, T., Masiero, J., Cutri, R., Dailey, J., Eisenhardt, P., McMillan, R., Wright, E., Walker, R., et al. (2011). Preliminary results from neowise: an enhancement to the wide-field infrared survey explorer for solar system science. *The Astrophysical Journal*, 731(1), 53.
- [135] Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., Prieto, C. A., Barkhouser, R., Bizyaev, D., Blank, B., Brunner, S., Burton, A., Carrera, R., et al. (2017). The apache point observatory galactic evolution experiment (apogee). *The Astronomical Journal*, 154(3), 94.

- [136] McDonough, E., Colin Hill, J., Ivanov, M. M., La Posta, A., & Toomey, M. W. (2024). Observational constraints on early dark energy. *International Journal of Modern Physics D*, 33(11), 2430003.
- [137] Melnik, R. & Willatzen, M. (2003). Bandstructures of conical quantum dots with wetting layers. *Nanotechnology*, 15(1), 1.
- [138] Ménard, B. & Bartelmann, M. (2002). Cosmological information from quasar-galaxy correlations induced by weak lensing. *Astronomy & Astrophysics*, 386(3), 784–795.
- [139] Ménard, B., Scranton, R., Schmidt, S., Morrison, C., Jeong, D., Budavari, T., & Rahman, M. (2013). Clustering-based redshift estimation: method and application to data. *arXiv preprint arXiv:1303.4722*.
- [140] Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239), 2.
- [141] Mishra-Sharma, S., Song, Y., & Thaler, J. (2024). Paperclip: Associating astronomical observations and natural language with multi-modal models. *arXiv preprint arXiv:2403.08851*.
- [142] Miville-Deschênes, M.-A. & Lagache, G. (2005). Iris: a new generation of iras maps. *The Astrophysical Journal Supplement Series*, 157(2), 302.
- [143] Mudur, N., Cuesta-Lazaro, C., & Finkbeiner, D. P. (2023). Cosmological field emulation and parameter inference with diffusion models. *arXiv preprint arXiv:2312.07534*.
- [144] Mudur, N. & Finkbeiner, D. P. (2022). Can denoising diffusion probabilistic models generate realistic astrophysical fields? *arXiv preprint arXiv:2211.12444*.
- [145] Multiphysics, C. (1998). Introduction to comsol multiphysics®. *COMSOL Multiphysics*, Burlington, MA, accessed Feb, 9(2018), 32.
- [146] Mustafa, M., Bard, D., Bhimji, W., Lukić, Z., Al-Rfou, R., & Kratochvil, J. M. (2019). Cosmogon: creating high-fidelity weak lensing convergence maps using generative adversarial networks. *Computational Astrophysics and Cosmology*, 6(1), 1–13.
- [147] National Agency for Finite Element Methods & Standards (Great Britain) (1990). *The Standard NAFEMS Benchmarks*. NAFEMS.

- [148] Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11), 2.
- [149] Nelson, D., Springel, V., Pillepich, A., Rodriguez-Gomez, V., Torrey, P., Genel, S., Vogelsberger, M., Pakmor, R., Marinacci, F., Weinberger, R., et al. (2019). The illustrious simulations: public data release. *Computational Astrophysics and Cosmology*, 6(1), 1–29.
- [150] Nguyen, N.-M., Schmidt, F., Tucci, B., Reinecke, M., & Kostić, A. (2024). How much information can be extracted from galaxy clustering at the field level? *arXiv preprint arXiv:2403.03220*.
- [151] Ni, B. & Buehler, M. J. (2024). Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. *Extreme Mechanics Letters*, 67, 102131.
- [152] Nichol, A. Q. & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning* (pp. 8162–8171): PMLR.
- [153] Ono, V., Park, C. F., Mudur, N., Ni, Y., Cuesta-Lazaro, C., & Villaescusa-Navarro, F. (2024). Debiasing with diffusion: Probabilistic reconstruction of dark matter fields from galaxies with camels. *arXiv preprint arXiv:2403.10648*.
- [154] OpenAI. Hello GPT-4o. Available online at: <https://openai.com/index/hello-gpt-4o/>.
- [155] OpenAI. o1. <https://openai.com/index/openai-o1-system-card/>.
- [156] Padmanabhan, N., Schlegel, D. J., Finkbeiner, D. P., Barentine, J., Blanton, M. R., Brewington, H. J., Gunn, J. E., Harvanek, M., Hogg, D. W., Ivezić, Ž., et al. (2008). An improved photometric calibration of the sloan digital sky survey imaging data. *The Astrophysical Journal*, 674(2), 1217.
- [157] Paganini, M., de Oliveira, L., & Nachman, B. (2018). Accelerating science with generative adversarial networks: an application to 3d particle showers in multilayer calorimeters. *Physical review letters*, 120(4), 042003.
- [158] Paillas, E., Cuesta-Lazaro, C., Percival, W. J., Nadathur, S., Cai, Y.-C., Yuan, S., Beutler, F., de Mattia, A., Eisenstein, D., Forero-Sanchez, D., Padilla, N., Pinon, M., Ruhlmann-Kleider,

- V., Sánchez, A. G., Valogiannis, G., & Zarrouk, P. (2023). Cosmological constraints from density-split clustering in the boss cmass galaxy sample.
- [159] Pan, H., Mudur, N., Taranto, W., Tikhonovskaya, M., Venugopalan, S., Bahri, Y., Brenner, M. P., & Kim, E.-A. (2024). Quantum many-body physics calculations with large language models. *arXiv preprint arXiv:2403.03154*.
- [160] pandas development team, T. (2020). pandas-dev/pandas: Pandas.
- [161] Papamakarios, G., Pavlakou, T., & Murray, I. (2017). Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30.
- [162] Pâris, I., Petitjean, P., Aubourg, É., Myers, A. D., Streblyanska, A., Lyke, B. W., Anderson, S. F., Armengaud, É., Bautista, J., Blanton, M. R., et al. (2018). The sloan digital sky survey quasar catalog: fourteenth data release. *Astronomy & Astrophysics*, 613, A51.
- [163] Parker, L., Lanusse, F., Golkar, S., Sarra, L., Cranmer, M., Bietti, A., Eickenberg, M., Krawezik, G., McCabe, M., Morel, R., et al. (2024). Astroclip: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4), 4990–5011.
- [164] Payeur, G., McDonough, E., & Brandenberger, R. (2025). Do observations prefer thawing quintessence?
- [165] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [166] Peek, J. & Graves, G. J. (2010). A correction to the standard galactic reddening map: passive galaxies as standard crayons. *The Astrophysical Journal*, 719(1), 415.
- [167] Peng, D. (2023). Langfun.
- [168] Pillepich, A., Springel, V., Nelson, D., Genel, S., Naiman, J., Pakmor, R., Hernquist, L., Torrey, P., Vogelsberger, M., Weinberger, R., et al. (2018). Simulating galaxy formation with the illustris model. *Monthly Notices of the Royal Astronomical Society*, 473(3), 4077–4106.
- [169] Poulin, V., Smith, T. L., & Karwal, T. (2023). The Ups and Downs of Early Dark Energy solutions to the Hubble tension: A review of models, hints and constraints circa 2023. *Phys. Dark Univ.*, 42, 101348.

- [170] Poulin, V., Smith, T. L., Karwal, T., & Kamionkowski, M. (2019). Early Dark Energy Can Resolve The Hubble Tension. *Phys. Rev. Lett.*, 122(22), 221301.
- [171] Prabhudesai, M., Ke, T.-W., Li, A. C., Pathak, D., & Fragkiadaki, K. (2023). Diffusion-tta: Test-time adaptation of discriminative models via generative feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- [172] Pramanick, S., Chellappa, R., & Venugopalan, S. (2024). Spika: A dataset for multimodal question answering on scientific papers. *arXiv preprint arXiv:2407.09413*.
- [173] Prusti, T., De Bruijne, J., Brown, A. G., Vallenari, A., Babusiaux, C., Bailer-Jones, C., Bastian, U., Biermann, M., Evans, D. W., Eyer, L., et al. (2016). The gaia mission. *Astronomy & astrophysics*, 595, A1.
- [174] Python Software Foundation (n.d.). difflib. Available online at: <https://docs.python.org/3/library/difflib.html>.
- [175] Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al. (2023). Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.
- [176] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
- [177] Régaldo-Saint Blancard, B., Allys, E., Auclair, C., Boulanger, F., Eickenberg, M., Levrier, F., Vacher, L., & Zhang, S. (2022). Generative models of multi-channel data from a single example—application to dust emission. *arXiv e-prints*, (pp. arXiv–2208).
- [178] Reid, B., Ho, S., Padmanabhan, N., Percival, W. J., Tinker, J., Tojeiro, R., White, M., Eisenstein, D. J., Maraston, C., Ross, A. J., et al. (2016). Sdss-iii baryon oscillation spectroscopic survey data release 12: galaxy target selection and large-scale structure catalogues. *Monthly Notices of the Royal Astronomical Society*, 455(2), 1553–1573.
- [179] Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- [180] Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

- [181] Remazeilles, M., Banday, A. J., Baccigalupi, C., Basak, S., Bonaldi, A., De Zotti, G., Delabrouille, J., Dickinson, C., Eriksen, H., Errard, J., et al. (2018). Exploring cosmic origins with core: B-mode component separation. *Journal of Cosmology and Astroparticle Physics*, 2018(04), 023.
- [182] Remy, B., Lanusse, F., Jeffrey, N., Starck, J.-L., Osato, K., & Schrabback, T. (2022). Probabilistic mass mapping with neural score estimation. *arXiv preprint arXiv:2201.05561*.
- [183] Rémy, B., Lanusse, F., Ramzi, Z., Liu, J., Jeffrey, N., & Starck, J.-L. (2020). Probabilistic mapping of dark matter by neural score matching. *arXiv preprint arXiv:2011.08271*.
- [184] Rezaei, S. K., Bailer-Jones, C., Hanson, R., & Fouesneau, M. (2017). Inferring the three-dimensional distribution of dust in the galaxy with a non-parametric method-preparing for gaia. *Astronomy & Astrophysics*, 598, A125.
- [185] Rezaie, M., Ross, A. J., Seo, H.-J., Mueller, E.-M., Percival, W. J., Merz, G., Katebi, R., Bunesco, R. C., Bautista, J., Brownstein, J. R., et al. (2021). Primordial non-gaussianity from the completed sdss-iv extended baryon oscillation spectroscopic survey-i: Catalogue preparation and systematic mitigation. *Monthly Notices of the Royal Astronomical Society*, 506(3), 3439–3454.
- [186] Riess, A. G., Yuan, W., Macri, L. M., Scolnic, D., Brout, D., Casertano, S., Jones, D. O., Murakami, Y., Anand, G. S., Breuval, L., et al. (2022). A comprehensive measurement of the local value of the hubble constant with 1 km s⁻¹ mpc⁻¹ uncertainty from the hubble space telescope and the shoes team. *The Astrophysical journal letters*, 934(1), L7.
- [187] Rogge, N. & Rasul, K. (2022). The annotated diffusion model.
- [188] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.
- [189] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241).: Springer.
- [190] Ross, A. J., Percival, W. J., Carnero, A., Zhao, G.-b., Manera, M., Raccanelli, A., Aubourg, E., Bizyaev, D., Brewington, H., Brinkmann, J., et al. (2013). The clustering of galaxies in the sdss-iii dr9 baryon oscillation spectroscopic survey: constraints on primordial non-gaussianity. *Monthly Notices of the Royal Astronomical Society*, 428(2), 1116–1127.

- [191] Rouhiainen, A., Gira, M., Münchmeyer, M., Lee, K., & Shiu, G. (2023). Super-resolution emulation of large cosmological fields with a 3d conditional diffusion model. *arXiv preprint arXiv:2311.05217*.
- [192] Rozet, F., Delaunoy, A., Miller, B., et al. (2021). LAMPE: Likelihood-free amortized posterior estimation.
- [193] Rybizki, J., Green, G., Rix, H.-W., El-Badry, K., Demleitner, M., Zari, E., Udalski, A., Smart, R., & Gould, A. (2020). A classifier for spurious astrometric solutions in gaia EDR3. VO resource provided by the GAVO Data Center.
- [194] Rybizki, J., Green, G. M., Rix, H.-W., El-Badry, K., Demleitner, M., Zari, E., Udalski, A., Smart, R. L., & Gould, A. (2022). A classifier for spurious astrometric solutions in gaia edr3. *Monthly Notices of the Royal Astronomical Society*, 510(2), 2597–2616.
- [195] Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al. (2024). Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.
- [196] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.
- [197] Schlafly, E., Finkbeiner, D., Jurić, M., Magnier, E., Burgett, W., Chambers, K., Grav, T., Hodapp, K., Kaiser, N., Kudritzki, R.-P., et al. (2012). Photometric calibration of the first 1.5 years of the pan-starrs1 survey. *The Astrophysical Journal*, 756(2), 158.
- [198] Schlafly, E., Meisner, A., Stutz, A., Kainulainen, J., Peek, J., Tchernyshyov, K., Rix, H.-W., Finkbeiner, D., Covey, K., Green, G., et al. (2016). The optical–infrared extinction curve and its variation in the milky way. *The Astrophysical Journal*, 821(2), 78.
- [199] Schlafly, E. F., Finkbeiner, D. P., Schlegel, D. J., Jurić, M., Ivezić, Ž., Gibson, R. R., Knapp, G. R., & Weaver, B. A. (2010). The blue tip of the stellar locus: Measuring reddening with the sloan digital sky survey. *The Astrophysical Journal*, 725(1), 1175.
- [200] Schlegel, D. J., Finkbeiner, D. P., & Davis, M. (1998a). Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *The Astrophysical Journal*, 500(2), 525.

- [201] Schlegel, D. J., Finkbeiner, D. P., & Davis, M. (1998b). Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *The Astrophysical Journal*, 500(2), 525.
- [202] Schmalzing, J. & Górski, K. M. (1998). Minkowski functionals used in the morphological analysis of cosmic microwave background anisotropy maps. *Monthly Notices of the Royal Astronomical Society*, 297(2), 355–365.
- [203] Schmalzing, J., Kerscher, M., & Buchert, T. (1995). Minkowski functionals in cosmology. *arXiv preprint astro-ph/9508154*.
- [204] Schöneberg, N., Abellán, G. F., Sánchez, A. P., Witte, S. J., Poulin, V., & Lesgourgues, J. (2022). The ho olympics: A fair ranking of proposed models. *Physics Reports*, 984, 1–55.
- [205] Sharma, D., Dai, B., & Seljak, U. (2024a). A comparative study of cosmological constraints from weak lensing using convolutional neural networks. *arXiv preprint arXiv:2403.03490*.
- [206] Sharma, D., Dai, B., Villaescusa-Navarro, F., & Seljak, U. (2024b). A field-level emulator for modeling baryonic effects across hydrodynamic simulations. *arXiv preprint arXiv:2401.15891*.
- [207] Shen, Z., Zhang, M., Zhao, H., Yi, S., & Li, H. (2021). Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3531–3539).
- [208] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.
- [209] Siegel, Z. S., Kapoor, S., Nagdir, N., Stroebel, B., & Narayanan, A. (2024). Core-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*.
- [210] Significant Gravititas (n.d.). AutoGPT.
- [211] Skrutskie, M., Cutri, R., Stiening, R., Weinberg, M., Schneider, S., Carpenter, J., Beichman, C., Capps, R., Chester, T., Elias, J., et al. (2006). The two micron all sky survey (2mass). *The Astronomical Journal*, 131(2), 1163.

- [212] Smith, M. J., Geach, J. E., Jackson, R. A., Arora, N., Stone, C., & Courteau, S. (2022). Realistic galaxy image simulation via score-based generative models. *Monthly Notices of the Royal Astronomical Society*, 511(2), 1808–1818.
- [213] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256–2265).: PMLR.
- [214] Song, Y., Shen, L., Xing, L., & Ermon, S. (2021a). Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*.
- [215] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- [216] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021b). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.
- [217] Strauss, M. A., Weinberg, D. H., Lupton, R. H., Narayanan, V. K., Annis, J., Bernardi, M., Blanton, M., Burles, S., Connolly, A., Dalcanton, J., et al. (2002). Spectroscopic target selection in the sloan digital sky survey: the main galaxy sample. *The Astronomical Journal*, 124(3), 1810.
- [218] Tamosiunas, A., Winther, H. A., Koyama, K., Bacon, D. J., Nichol, R. C., & Mawdsley, B. (2021). Investigating cosmological gan emulators using latent space interpolation. *Monthly Notices of the Royal Astronomical Society*, 506(2), 3049–3067.
- [219] Tange, O. (2018). *GNU Parallel 2018*. Ole Tange.
- [220] Theis, L., Oord, A. v. d., & Bethge, M. (2015). A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.
- [221] Tian, C. & Zhang, Y. (2024). Optimizing collaboration of llm based agents for finite element analysis. *arXiv preprint arXiv:2408.13406*.
- [222] Tian, M., Gao, L., Zhang, S. D., Chen, X., Fan, C., Guo, X., Haas, R., Ji, P., Krongchon, K., Li, Y., et al. (2024). Scicode: A research coding benchmark curated by scientists. *arXiv preprint arXiv:2407.13168*.

- [223] Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature*, 625(7995), 476–482.
- [224] Turon, C., Meynadier, F., Arenou, F., Lindegren, L., & Bastian, U. (2010). Basic principles of scanning space astrometry. *European Astronomical Society Publications Series*, 45, 109–114.
- [225] Valogiannis, G. & Dvorkin, C. (2022). Going beyond the galaxy power spectrum: An analysis of boss data with wavelet scattering transforms. *Physical Review D*, 106(10), 103509.
- [226] Valogiannis, G., Yuan, S., & Dvorkin, C. (2023). Precise cosmological constraints from boss galaxy clustering with a simulation-based emulator of the wavelet scattering transform. *arXiv preprint arXiv:2310.16116*.
- [227] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [228] Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Spergel, D. N., Li, Y., Wandelt, B., Nicola, A., Thiele, L., Hassan, S., Matilla, J. M. Z., et al. (2021a). Multifield cosmology with artificial intelligence. *arXiv preprint arXiv:2109.09747*.
- [229] Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Spergel, D. N., Somerville, R. S., Dave, R., Pillepich, A., Hernquist, L., Nelson, D., Torrey, P., et al. (2021b). The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1), 71.
- [230] Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., Thiele, L., Dave, R., Narayanan, D., Nicola, A., Li, Y., Villanueva-Domingo, P., Wandelt, B., et al. (2022). The camels multifield data set: Learning the universe’s fundamental parameters with artificial intelligence. *The Astrophysical Journal Supplement Series*, 259(2), 61.
- [231] Villanueva-Domingo, P. & Villaescusa-Navarro, F. (2022). Learning cosmology and clustering with cosmic graphs. *The Astrophysical Journal*, 937(2), 115.
- [232] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J.,

- Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272.
- [233] Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2024a). Examining the potential and pitfalls of chatgpt in science and engineering problem-solving. In *Frontiers in Education*, volume 8 (pp. 1330486).: Frontiers Media SA.
- [234] Wang, X., Chen, Y., Yuan, L., Zhang, Y., Li, Y., Peng, H., & Ji, H. (2024b). Executable code actions elicit better llm agents. *arXiv preprint arXiv:2402.01030*.
- [235] Waters, C., Magnier, E., Price, P., Chambers, K., Burgett, W., Draper, P., Flewelling, H., Hodapp, K., Huber, M., Jedicke, R., et al. (2020). Pan-starrs pixel processing: detrending, warping, stacking. *The Astrophysical Journal Supplement Series*, 251(1), 4.
- [236] Wildberger, J., Dax, M., Buchholz, S., Green, S., Macke, J. H., & Schölkopf, B. (2024). Flow matching for scalable simulation-based inference. *Advances in Neural Information Processing Systems*, 36.
- [237] Wright, E. L., Eisenhardt, P. R., Mainzer, A. K., Ressler, M. E., Cutri, R. M., Jarrett, T., Kirkpatrick, J. D., Padgett, D., McMillan, R. S., Skrutskie, M., et al. (2010). The wide-field infrared survey explorer (wise): mission description and initial on-orbit performance. *The Astronomical Journal*, 140(6), 1868.
- [238] Wu, Y. & He, K. (2018). Group normalization. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).
- [239] Yahata, K., Yonehara, A., Suto, Y., Turner, E. L., Broadhurst, T., & Finkbeiner, D. P. (2007). The effect of fir emission from sdss galaxies on the sfd galactic extinction map. *Publications of the Astronomical Society of Japan*, 59(1), 205–219.
- [240] Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., & Press, O. (2024a). Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793*.
- [241] Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., & Anandkumar, A. (2024b). Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*, 36.

- [242] Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., Kiraly, A., Ahmed, F., Hormozdiari, F., Jaroensri, T., et al. (2024c). Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.
- [243] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- [244] Yasuda, N., Fukugita, M., & Schneider, D. P. (2007). Spatial variations of galaxy number counts in the sloan digital sky survey. ii. test of galactic extinction in high-extinction regions. *The Astronomical Journal*, 134(2), 698.
- [245] Zagoruyko, S. & Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.
- [246] Zhou, R., Guy, J., Koposov, S. E., Schlafly, E. F., Schlegel, D., Aguilar, J., Ahlen, S., Bailey, S., Bianchi, D., Brooks, D., et al. (2024). Stellar reddening map from desi imaging and spectroscopy. *arXiv preprint arXiv:2409.05140*.
- [247] Zonca, A., Singer, L., Lenz, D., Reinecke, M., Rosset, C., Hivon, E., & Gorski, K. (2019). healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in python. *Journal of Open Source Software*, 4(35), 1298.



THIS THESIS WAS TYPESET using $\text{\LaTeX}$, originally developed by Leslie Lamport and based on Donald Knuth's $\text{\TeX}$. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under [CC BY-NC-ND 3.0](#). A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.