DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery

Citation

Pyzer-Knapp, Edward O., Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik. 2015. "What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery." Annual Review of Materials Research 45 (1) (July): 195–216. doi:10.1146/annurev-matsci-070214-020823.

Published Version

doi:10.1146/annurev-matsci-070214-020823

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:25977966

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

What is High Throughput Virtual Screening? A Perspective from Organic Materials Discovery

Edward O. Pyzer-Knapp, Changwon Suh,Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik*

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA, 02143

Annual Reviews of Materials Research
2014. AA:1–25

This article's doi:
10.1146/((please add article doi))

Copyright © 2014 by Annual Reviews.
All rights reserved

Keywords

Computational materials design, Big Data

Abstract

A philosophy for defining what constitutes a virtual high-throughput screen is discussed, and the choices which influence decisions at each stage of the 'computational funnel' are investigated, including an in-depth discussion of the generation of molecular libraries. Additionally advice on the storing, analysis and visualization of data is given, based upon extensive experience in the group.

Contents

1. Introduction

As a society we categorize our history by either the prevalent materials (e.g. Bronze Age, Iron Age, etc) or by the groundbreaking processes (e.g. the Industrial Revolution) related to their manufacture. It would not be unreasonable to say that we are now in the age of Materials Science - an age best categorized by the cornucopia of available materials made possible by a scientific method for discovery. The path to the present day, however, has not been simple, or well defined. Many of the most significant materials discoveries have been made not by rational design, but instead are a product of happenstance, where the stars aligned and the right person was in the right place at the right time. Perkins's mauve, vulcanised rubber, and teflon are famous examples of this process. This should not be surprising, since the size of chemical space - recently estimated at $> 10^{60}$ molecules (1) - makes any kind of rational global search challenging in the extreme.

Fortunately, global exploration is rarely required - since we often have a good idea of the local area of chemical space in which we would like to explore. This reduces the size of the exploration from the order of 10^{60} molecules to somewhere on the order of 10^{6}. Whilst this is still far too many molecules for even the most advanced experimental screening techniques to consider, the massively parallel nature of the materials screening problem coupled with recent advances in computer architecture and distribution techniques have born the concept of a virtual high throughput screen - in which large libraries of molecules are analyzed using theoretical techniques, and reduced to a small set of promising leads for

experimental chemists to follow up on. Indeed, some have postulated that because of this, we are on the cusp of a *Golden Age* of Materials Discovery. (2) We discuss the philosophy that defines a high throughput screen, focusing on key areas such as size of library, hierarchical techniques, and analysis methods. It should be noted that these techniques solely focus on the discovery of a new material, and not the commercialization process, which is typically much longer.

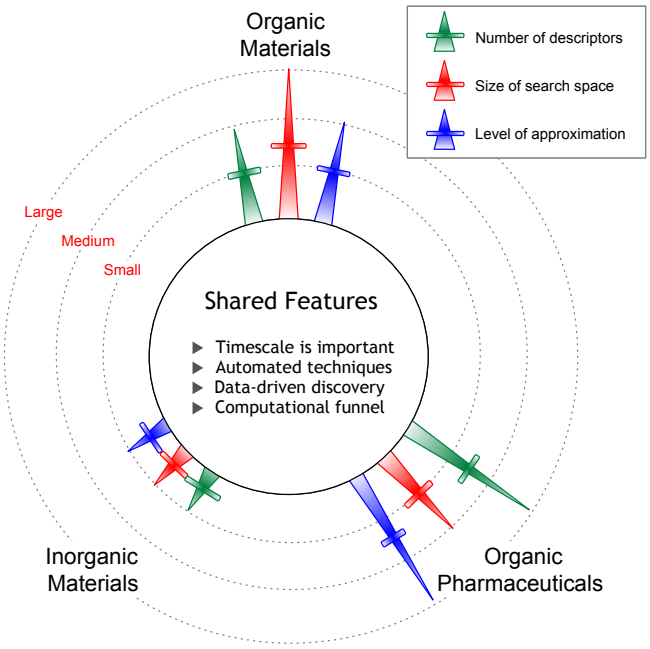

Figure 1

High throughput screening techniques from many different areas share a core philosophy, despite their computational needs being of different magnitudes.

2. The Philosophy Behind High Throughput Virtual Screening

Whilst high-throughput virtual screening has existed (particularly in the pharmaceutical sciences) for a while, its definition is somewhat intractably tied to the capability of the state of the art hardware at the time of calculation. We therefore propose that this field is better defined by the philosophy of the calculations undertaken rather than the number, or intensity.

2.1. Context

Whilst the context of the calculations and chemical space differ between inorganic materials, organic materials, and organic pharmaceutical chemistry, high throughput screens in all of these areas share the same underlying philosophy. Together, these statements form a

Four Philosophies of High Throughput Virtual Screening

1. Timescale is important; time critical techniques require high-throughput solutions
2. Automated techniques; high-throughput approaches require automation
3. Data-driven discovery; Trends in the data are often as important as the data itself
4. Computational Funnels; A funnel like approach, in which only promising molecules are exposed to expensive calculation methods, allows efficient deployment of computation

definition of a high-throughput screen. We expand upon each of these, before reviewing their application in the literature as well as from our own research experiences.

2.2. Time critical problems require high throughput solutions

Research moves at a variety of speeds. Whether it is due to pressure from competitors, financial motives, or social pressures, it is often desirable to use a high throughput technique to quickly focus on one particularly promising part of chemical space. Additionally some problems are intrinsically time critical. Our supplies of fossil fuels are finite, and rapidly dwindling; however our consumption continues to increase year on year. New technologies must be both developed and implemented soon to allow us to continue to feed our energy needs. Clearly the timescale of studies such as these is of sufficient importance that it must influence the design of the investigation. It is in this area that high throughput methodologies deliver greatest value.

The timescale, whether required or desired, is one of the factors which we believe defines a high throughput screen.

2.3. High-throughput approaches require automation

By its very definition, it is at the least exceptionally challenging - and often simply impossible - to manually perform a high throughput screen. The generation, storing, and querying of the large volume of data require some degree of automation to be efficient, and these aspects will each be discussed in more detail in the upcoming sections. The increased use of combinatorial techniques adapted from life sciences for the generation of libraries has allowed the creation of initial libraries of millions of candidate molecules; leaving a manual approach far beyond the ability of even the most fervent of investigators.

High throughput screens require automation, especially in the early stages

2.4. Computational Funnels allow efficient deployment of computation

Frequently, the calculation of the property of interest in a screen is intrinsically too expensive for mass deployment over all potential molecules thus potentially placing it outside the realms of a high throughput approach. One approach to avoiding this pitfall is to employ a computational funnel. Each level of the funnel represents a calculation with well defined error-bounds, and at each level structures are ruled out based upon selection criteria defined by the error-bounds. Since each level is progressively more computationally intense, only the

molecules most likely to be of interest are calculated with the most expensive methods, with each new level affording additional information about the molecule. The final 'test' is an experimental fabrication of a device as close as possible to the expected running conditions, and since this is both slow and expensive, the fewer number of candidate molecules which reach this stage, the better.

Following a computational funnel allows us to focus computational effort on promising molecules

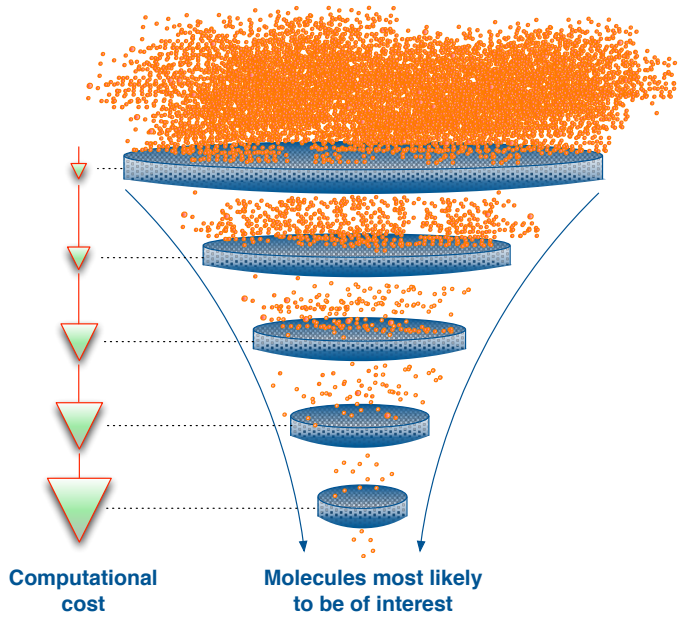

Figure 2

A computational funnel scheme. The increasingly strict filtering criteria eliminates many molecules that are not of interest and identifies the top performing candidates in a virtual library.

3. Molecular Libraries

3.1. General considerations

Every project that aims at a high-throughput screening of a section of molecular space must begin by selecting the candidate molecules to be investigated. This process is crucial, since it puts boundaries to the possible outcomes from the start; the successful candidate can only come from the initial library, or from its successive rounds of growth.

The generation of libraries is particularly important in exploratory research, when it is not known what type of molecule will solve a given problem, and somewhat less so when one one is performing a thorough crib through some fixed finite subsection of space. That is, generating novel lead backbones is a harder challenge than pursuing a complete set of substitution patterns through side chain enumeration

In molecular space, it is particularly difficult to perform systematic explorations: we

lack any predefined magnitudes to survey chemical space, so that we can move systematically along them and create a search grid or more sophisticated search pattern. Molecular structure obeys a large and complex set of rules that so far defies systematic exploration.

Despite the obvious, and also subtle, dissimilarities with drug discovery in the pharma industry, high throughput screening of organic materials has much to learn from its older and larger sibling. In the field of drug discovery it is common to explore known chemical space for novel applications rather than try to create libraries de novo. (3), or at least, perform variations starting on a known drug, after the maxim: 'the most fruitful basis for the discovery of a new drug is to start with an old drug'. (4) This is due to a combination of reasons, such as the large capital costs of developing new drugs, the high cost of late-stage failures or the fact that structure-biological activity relationships are usually much more obscure than structure-property relationship in the less mature area of organic materials screening.

Given the intrinsic impossibility of enumerating all the molecules in all but the smallest sections of molecular space, much effort is being put into methods to avoid explicit enumeration in high throughput screening, with approaches such as optimizing potentials to make a rugged region of chemical space flatter, (5) alchemical transformation on a given backbone (6), generating aromatic rings that electronically equivalent to a reference; (7) stochastic generation of derivatives; (8), recursive substructure searches (9) or morphing of starting molecules of interest. (10)

However, more often than not, high-throughput-screening projects rely on explicitly enumerated libraries. Led by the need for diverse drug-like molecules, many examples of computer-generated molecular libraries have spawned in recent years, oriented mostly at compounds with potential biological activity. The chemical universe generated databases aim at exhaustiveness: GDB-11 with 14 million molecules containing up to 11 atoms of C, N, O, and F; (11) GDB-13 with under one billion molecules containing up to 13 atoms of C, N, O, S, and Cl; (12) GDB-17 with 166 billion molecules; (13). Many other more oriented databases exist aiming at exploring smaller subsets of larger molecules, such as tetrapyrrole macrocycles; (14), natural product-like virtual libraries through recursive atom-based enumeration (15) or the recently reported ZINClick database of 16 million triazines (16)

3.2. Molecular diversity

Molecular diversity, is a vital concept in screening of molecular libraries which material screening has also inherited from pharma. Research in virtual molecular libraries has attracted much attention into trying to maximize molecular diversity within a given library. (17) The aim of diverse libraries is to do a homogeneous multidimensional crib of molecular space, where little effort is spent exploring the neighborhood of areas already represented in the library.

The key issue however is that we lack absolute axes along which to survey molecular space (18) and thus, we do not have universal metrics to assess similarity: a single iso-electronic alchemical substitution can completely disrupt the electronic structure of a molecule, whereas large variations in side chains can be relatively harmless to optical properties. Thus, what similarity metric to use, and ultimately how diverse a library is, is more an ad hoc decision based on chemical intuition than an intrinsic property of a library. (19, 20)

3.3. Generation of custom-made libraries

As mentioned above, the connections between molecular structure, electronic structure and device properties in organic electronic materials are more straightforward than the structure-activity relationships in drug discovery, where even knowing the macromolecular target hardly limits the choice of potential scaffolds. Since the cpu cost of computational methods in organic materials is quite high (see below), it is important to have as high a hit-ratio has possible, and thus it is often more rewarding to generate custom libraries to trawl promising regions of molecular space, rather than using generic predefined ones. Many different codes have been reported for the assembly of virtual molecular libraries, oriented essentially towards lead discovery and optimization in pharma. (21, 22, 23, 24, 25, 26, 27, 28, 29, 30).

Library creation implies generating some or all the possible combinations of a given set of fragments in a combinatorial way according to a set of rules. The fragments and the rules may borrow from experimentally feasible combinatorial synthesis schemes, or they may just be arbitrary schemes to explore chemical space with only indirect connection to chemical synthesis.

In addition, it is necessary to set a limit when the growth of the molecule will come to and end. This termination procedure will define the maximum size of a given molecule and thus set a ceiling for computational cost, helping estimate the resources needed. In the most trivial case, a one-off combinatorial linkage, the maximum size is easily estimated with the maximum size of the two fragments to be joined. In more complex cases with variable growth steps, stopping points can be fixed at given round of growth, maximum atom or electron count or maximum molecular mass.

The challenge of selecting a slice of molecular space cannot be overstated and deep chemical knowledge can be leveraged to generate libraries that explore novel regions and exploit promising ones while keeping the number of false positives as low as possible.

In addition, the fundamental target is to produce a library of molecules that fulfill some property requirement, that are also accessible, *i.e.* that are possible to synthesize, and ultimately represent good value for investment - both time and economic. These factors represent soft constraints, that change not only between projects but ultimately within a project: a more challenging synthesis can be pursued for a higher payoff. (31) It is important to address these contstraints at the earliest point - the construction of the molecular library. Substitutions patterns in the fragments, and mode of growth are key items here: a molecule is more likely to be synthesizable if the substitution pattern is the same in identical positions within a moiety.

The next challenge of molecular library generation is to encode these soft synthesis constraints into algorithmic molecular growth rules, so that as little time as possible is wasted synthetically inaccesible regions of molecular space. Chemical intuition can be leveraged in the generation of libraries to maximize the chances of discovering molecules that are both synthetically accessible and useful for a given application. For instance, using constrains based on synthetic accessibility, O'Boyle et al. reduced a set from potentially 800 million combinations to a mere 60 thousand in a search for organic photovoltaic materials. (32, 33)

There is, however, a tradeoff to encoding these soft constraints into hard algorithmic rules: since only what is in the library gets to be screened, one risks leaving out of the process molecules that are harder to make but perhaps have game-changing properties: the high-risk high-reward scenario. It would be desirable then, to assess synthetic availability just like any other property along its own scale, and molecules can be judged globally.

Despite large efforts to automatically assess synthetic availability, (34, 35, 36, 37, 38) we have yet to reach the point where these synthetic constraints can be extracted away from the library generation process and passed on to another score.

The computational efficiency can also be improved by hard coding some constraints into the library generation software, by prohibiting and filtering out molecules carrying certain functional groups that may arise during the molecular generation procedure and that are either fundamentally unstable for the foreseen application or detrimental to the property of interest.

In the following paragraphs three specific examples, currently being pursued by our own research group, of custom-generated libraries are reported, in the areas on organic photovoltaics, small molecule organic based flow batteries and organic light emitting diodes.

3.4. Organic photovoltaics

Given humanity's ever-increasing appetite for energy and the obvious drawbacks of conventional non-renewable energy sources, developing materials to harvest solar energy has been one of the targets of high-throughput virtual screening. The Clean Energy Project (CEP) is an effort for the the discovery of the next generation of plastic solar cell materials (39). So far, over 3 million molecules have been generated and a total of 300 million DFT calculations have been performed to identify low-cost, high-efficiency organic photovoltaics.

All the molecules in the CEP library were grown from an initial selection of 26 fragments. (40) The main source for this initial library is chemical intuition from experimental collaborators. This expertise was leveraged not only in the fragment selection, but also in the definition of the positions through which the fragments can be joined.

The strategy followed in this case included two possible fragment combinations. The first one, "linking" created a single covalent bond between the two fragments undergoing the reaction. The second one, "fusion" was used when both fragments include rings. This reaction made a final molecule where a fused ring is created with the two original fragments sharing a covalent bond, as can be seen in **Figure 3**. It is important to note that this second process requires the loss of two carbon atoms, so it is not related in any way to a chemical process, being a purely computational construction. The cutoff for growth was number of generations, with molecules all the way to tetramers (combination of four original fragments), and a total number of molecules to be screened of over 3 million.

3.5. Organic-based flow batteries

Renewable energy sources such as sunlight or wind have a more unpredictable output than traditional non-renewable ones, and thus a shift towards greener energy sources will require developments in energy storage to compensate the highs and lows in energy production. Flow batteries are a promising response, and Aziz *et al.* have proposed the use of anthraquinone derivatives as electrolytes in flow batteries for massive storing of electrical energy. (41) The use of organic redox species opens the doors to sustainable sourcing the electrolytes. The choice of anthraquinone redox molecule was helped by a combinatorial screening involving R-group enumeration in multiple quinone backbones.

The goal of molecular generation in this project was to pursue of a complete set of substituion patterns through R-group enumeration. The molecular frameworks were set from the start (benzo-, naptho- and anthraquinone), and the full combinatorial space of substitution with one or multiple instances of each functional group were explored. These

patterns were applied to several R-groups including sulphonate, hydroxy and phosphonate. The nature of the side groups, mostly their electron-withdrawing or electron-donating and their polarity, tune the two key chemical properties for these materials: redox potential and solubility.

The "linking" procedure described before is used to generate all the possible combinations of core and R-group. Libraries that explore full R-group enumeration can be very sizeable, since they grow factorially with the number of positions: in this case, all the possible 1,2; 2,3 and 1,4 quinones are possible for a given ring backbone, and in each of this, every remaining carbon atom in the quinone ring can bear a functional group. See Figure 3. A key issue to take into account is how many different R-groups to be combined in a given molecule. For a quinone with 8 available C-H positions and a single R-group, the number of posible molecules (excluding symmetry) is 8!. With the inclusion of a second R-group, $8! \cdot 7!$, and with a third, $8! \cdot 7! \cdot 6!$. The number becomes astronomical even for a small number of different R-groups. A very early cutoff has to be chosen regarding the maximum amount of different groups in a given molecule. In this particular case, that number was limited to 2.

There is no need to select a termination strategy since the maximum size is determined by the size of the core plus that of the substituents. In a first library, (41) the substitution pattern explored was that of singly oxidized quinones either singly or fully substituted with one R group within a list of 14. This was widened after to benzo-, naphto- and anthraquinones bearing two C=O groups and any substitution pattern with any instances of a single R-group (with a total of 3037 unique substitution patterns).

3.6. Blue organic light emitting material

Recent developments in thermally-assisted, delayed fluorescence (TADF) have opened the door to novel clases of organic light-emitting diodes (OLED). (42, 43) These novel molecules exhibit low enough splitting between their lowest singlet and triplet states that efficient thermal re-population of the emissive singlet from the dark triplet is possible. Low-splitting excitations correspond to charge transfer states, and thus the basic TADF OLED must include electron donor and electron acceptor moieties, with some linker breaking the π conjugation.

TADF molecules molecules have to satisfy a very specific Donor-(Bridge)-Acceptor structure. As described above, it is advisable to encode as much chemical knowledge as possible into the library generation process, to avoid spending valuable screening time in areas of chemical space that will be barren by necessity. One can picture the inefficiency of a scenario where fragments are combined without restrictions and only an analysis a posteriori would lead to the desired configuration.

Three successive strategies have been applied in the fragment selection. Initially, fragments were selected that have been present in the OLED literature. In a second effort, a new set of fragments not related with the OLED literature were selected, underwent a screen to trim undesirable candidates (in this case, according to the HOMO and LUMO positions and optical properties) and in a further step, to facilitate the synthesis of the final molecule synthetic availability of each fragment was confirmed in the literature.

The third and final strategy, included a random generation of fragments; creating one-, two- and three-ring heterocylces with various amounts of nitrogen, oxygen and sulfur and a pre-screen for electronic properties. This a big-risk big-reward scenario, where completely

new fragments not explored or synthesized before are studied.

As is the case with the organic flow battery project, only the "linking" described before was used. To fulfill a Donor-Acceptor strategy, the Donor and Acceptor space was first expanded with combinations of each of them with Bridge fragments in a symmetrical fashion. This symmetry constrain, i.e. that analogous positions in a molecule grow in a identical way, is paramount for increasing the synthetic availability of molecules, while restraining the combinatorial nature of the growth and reducing the computational cost of screening by orders of magnitude. In a final step the Donor and Acceptor parts are combined with each other. For a small example, see **Figure 3**. A maximum molecular weight is determined to limit the size of the final molecules generated.

3.7. Other considerations

We have very broadly explained three different strategies to achieve the generation of a set of molecules to be used for screening. There are a few ideas to be considered to further improve this process. We have seen the use of the concept of symmetry to guide synthesis. There are other ways to improve this process. We can forbid the creation of certain covalent bonds that may be very difficult or impossible to make. Similarly, if there are factors known about the physical processing of the final product (e.g. evaporation) it may be desirable to exclude molecules with certain limiting physical properties (e.g. the molecular weight) from the molecular library.

The most crucial point, at least conceptually, is the fact that once the fragments and the way they combine is selected, we have limited ourselves to a very small area of the molecular space. No molecule outside that area will be screened. The decision made is a compromise between including a big part of the molecular space and keeping the computational expenses tractable. We should be able to establish a feedback loop to rethink the generated set with new fragments or combinatorial strategies when new information about the chemical space is available.

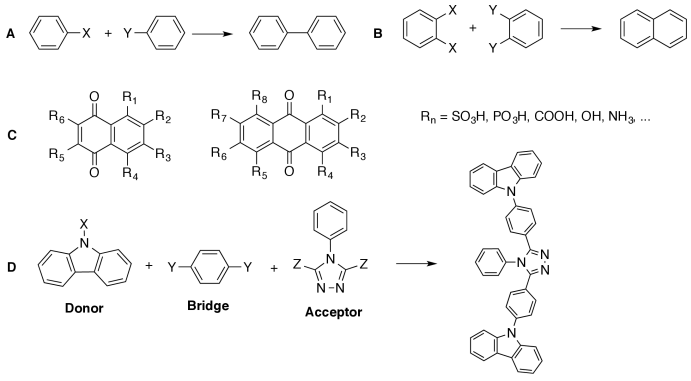

Figure 3

Reactions and combinations in virtual library enumeration. 'A' shows the linking procedure, used in the blue OLED project, and 'B' the fusion procedure additionally utilized in the Clean Energy Project. 'C' shows the enumeration of the different substitution positions considered in the organic-based flow battery project, and 'D' shows how a donor, bridge, and acceptor molecule combine to give a potential blue OLED material.

4. On the Theoretical Calculation of Materials Properties

The selection of simulation tools for high-throughput materials screening must be guided by clearly defined objectives in terms of the physical and chemical properties that are desired or necessary for the target technology. In other words, what are the optimization parameters and what are the constraints? In order to answer this, it is often necessary to ask further questions about the desired physical properties of the material. For instance, under what conditions will the material be required to remain stable, will processing conditions place limits on molecular weight or solubility, and what properties will be associated with a cheap and safe material? Is the target property thermodynamically or kinetically limited? Partnerships with industry may be useful in identifying the right constraints since synthesis or processing considerations may differ significantly between academic and industrial laboratory settings. The answers to these questions may guide the selection of the list of properties to screen for. For instance, a screen for battery electrolyte solvents might focus on a desired electrochemical stability window, melting and boiling point temperatures, viscosity, dielectric constant, Li-ion conductivity, and electronic conductivity ranges. Of these properties, only those that can be estimated with suitably cheap computational models can be chosen as initial screening criteria. In the realm of organic materials, this typically means that only single molecule properties may be calculated, and solvent or solid-state effects, must be estimated through an empirical model. In the example of electrolytes, the electrochemical stability window is an easy target property for quantum chemical simulations. Studies on a small family of related compounds will often guide the choices of what methodology is approriate at different levels of a staged screen. We cannot emphasize enough that domain-specific knowledge must guide the selection of both search space and computational tools since judicious choices will vastly improve both the efficiency of the expended computer time and the viability of top performing candidates.

Computational methods for screening material properties might broadly be broken into the following simulation categories: quantum mechanical methods (semi-empirical, density functional theory, or wave function theory), classical force-field based methods, and data-driven paradigms (encompassing quantitative structure property relationships, genetic algorithms and machine learning approaches). By virtue of targeting high-throughput computation, computational cost sharply limits the available computational techniques. However, in screening we are only interested in the ranking of candidates, and therefore systematic shifts between computational and experimental results need not be concerning, as long as the faster method does not introduce so much error that we are unable to correctly identify the top hundred or so candidates. (44)

We demonstrate how the choice of property, and the wider context of the screen, can influence the methods used by looking in detail at three areas:

1. Inorganic Li-ion batteries (e.g. The Materials Project)
2. Organic Photovoltaics (e.g. The Clean Energy Project)
3. Organic-Based Flow Batteries

4.1. The Materials Project

The Materials Project is a high-throughput effort focused upon traversal of the inorganic materials space in search of novel battery materials run by Professor Gerbrand Ceder (MIT) and Dr. Kristin Persson (Lawrence Berkeley National Laboratory). To date they have computed relevant properties of over 80,000 materials and screened 25,000 of these for use as

potential Li-ion batteries. This has so far used over 15 million CPU-hours of computational time at the National Energy Research Scientific Computing Center (NERSC).

Whilst the Materials Project mainly utilizes 'traditional' compute sources (large super computer clusters) the results are distributed through a web portal, (45) and can be searched and analyzed using a custom python module *PyMatGen*. (46)

Since the Materials Project is interested in electronic properties, a method based upon quantum mechanical principles is required. The high-throughput DFT computations were performed using the Vienna software package (VASP) (47), using the PAW pseudopotentials (48) and the Generalized Gradient Approximation (GGA) (49). GGA was chosen since it represents a good compromise between speed and accuracy. To compensate for the known errors in the model due to electron self-interaction energy, calculations were performed within the DFT+U framework. (50) Ceder *et al.* note that the use of hybrid functional - which reduce the effects of the self interaction energy term - could increase the accuracy of the calculation and remove the need for operating within the DFT + U framework. These were, however, considered too expensive for use in a high throughput screen. (51)

It is sometimes the case that one particular method of calculation will have systematic failings for particular classes of molecule. The Materials Project aims to avoid any such biases within its data set by classifying materials into different classes, each with their own specific battery of associated calculations. The results from these calculations are then unified over the global population using a set of reference reactions to connect results from different methods. (52)

4.2. The Harvard Clean Energy Project

The Harvard Clean Energy Project is a search for molecules with the potential to be employed in organic photovoltaic devices. (39, 53) It is unique amongst materials high-throughput screening projects for the scale at which it utilizes distributed computing through the World Community Grid. (54) It has been estimated that the computing power that this affords the project is in the order of a fully utilized 6-7,000 core cluster, and the amount of harvested CPU time can be seen in **Figure 4**.

With such a large amount of computing power available, a slightly different approach was taken to the Materials Project. In place of one fast functional (GGA), the Clean Energy Project calculates properties of molecules with a range of different functional both to reduce systematic errors that occur in particular functionals, and also to investigate how the choice of functional affects the values computed. The basic work-flow consists of two types of job; an optimization of the molecular geometry, and a single point calculation on that optimized geometry. Even with the vast resources available, optimizations using all the different functionals was not considered a good deployment of computational power, and so the geometry as optimized using the BP86 functional was used for all single point calculations.

For solar cell performance, the electronic structure itself is not the property of interest; rather it is how the electronic structure interacts with photons of sunlight that provides the true ranking metric. For this, the Scharber model of the power conversion efficiency (PCE) (55) was used to rank molecules. The Scharber model is a specialized version of the Shockley–Queisser model (56) for OPVs and is based upon the energies of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO)

- which are both easily calculable properties.

Due to its large size (over 3 million molecules, and over 300 million quantum-chemical calculations), the Clean Energy Project Database (57) provides an ideal test-bed for data-driven approaches such as cheminfomatics and machine learning. These methods represent the potential for quickly and rigourously developing Qualitative Structure Property Relationship (QSPR)-type models from existing data, which can then be used to focus calculation effort upon promising areas of chemical space. Olivares-Amaya *et al.* have used cheminformatics descriptors to this effect, calculating the current–voltage properties of over 2.5 million molecules using linear regression descriptor models. (40)

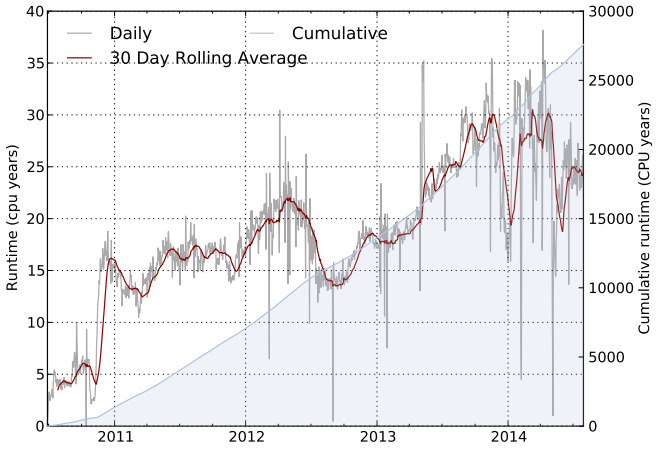

Figure 4

The amount of CPU harvested over the course of the Clean Energy Project calculated as 1-CPU runtime equivalent.

4.3. Organic-Based Flow Batteries

The search for a metal free flow battery by Aziz *et al.* (41) represents a good example of how a high throughput virtual screen can complement an experimental study, improving the efficiency of optimizing the molecule within a chemical space.

An essential component of this new generation of aqueous redox flow battery is a unique quinone molecule, the 9,10-anthraquinone-2,7-disulphonic acid (AQDS). The small, electroactive, and water-soluble AQDS molecule that is used at the negative side of a flow battery was identified using quantum chemical calculations on a pool of approximately 10,000 candidate molecules. Computational studies were focused on to investigate how changing the substitution patterns of AQDS would affect two key properties:

1. The redox potential, E^0, of the quinone-hydroquinone couples
2. The solvation free energy, G^0_{solv}, of quinones in water

For a high-throughput search such as this, where the generation of three-dimensional structures (conformers) are an important component of calculation workflow. As with any calculation, there exists a trade-off between computational cost of a calculation and

accuracy. Since the generation of many conformers represents a more complete exploration of the energy landscape of a molecule, an additional criteria is introduced: the *completeness* of the search.

With a high-throughput screening, the number of molecules to be processed dictates the amount of time that the argorithm can spend on each molecule. Conformer generation algorithms can be generally split into two categories:

1. Physically motivated generators (e.g. a low-mode generator (58))
2. Rule based generators (e.g. Corina (59))

The most reliable way to have an acceptably complete search (ignoring the trivial case of simply varying all torsion angles and calculating the energy - which is simply too costly to use in anything other than the simplest of cases) is to use a physically motivated generator, such as the low-mode conformer generator implemented in MacroModel. (60) These techniques have shown good results in the many different circumstances, but are also slow to run. In general, the rule based conformer generators are much faster than their physically based counterparts. The price that comes with that speed increase is that, due to their very nature, they are only applicable to molecules which are similar to those which were used to develop the rules on which they are based. Since the molecules within Aziz *et al.*'s search were all similar; it was determined that a rule based search was acceptable, and starting geometries were improved by minimization using the Dreiding force field, (61) from which a low energy selection were then re-minimized using density functional theory (GGA/PBE) - which, as discussed in the previous example, represents a good compromise between speed and accuracy.

This project continued to use the concept of the computational funnel with its approach to calculating E^0 and G^0_{solv} - the properties that are of primary interest for the development of flow batteries utilizing water-soluble, electroactive molecules.

5. Considerations for High Throughput Virtual Screening

5.1. Deployment of a High Throughput Virtual Screen

Having decided that high throughput virtual screening is the correct solution to a particular scientific problem, focus now shifts onto the deployment of the calculations which will produce the data of interest. The university, or company, computing cluster is the traditional locale for these computations. The compute capability across these types of machine is typically very isotropic and reliable, with a file-system that will allow the running of a large number of intense calculations at the same time. The one major downside to using these resources is that they are typically shared amongst many users, limiting the time any one user can get for their own projects. One other alternative option that has gained traction recently is the use of distributed computing. This uses idle time on the computers of volunteers to achieve a facsimile of the compute cluster by borrowing compute time from these transient 'nodes'. Popularized by the 'Folding at Home' (62) and 'SETI at Home' (63) projects, this approach has yielded good results for a wide variety of projects, especially through IBM's World Community Grid initiative. As demonstrated in Figure 4, this can result in significant amounts of CPU resource - invaluable to a high-throughput virtual screen. Of course there are downsides to this approach, which mainly result from the fact that calculations are being performed upon computers whose main purpose is not simply compute, but so long as the project requirements are not too strenuous this remains a valid

approach - indeed we are starting to see universities utilise it for their own research. (64)

5.2. Dealing with Data from High Throughput Screens

Data produced as a product of a high-throughput screen produces a series of challenges, which may be unfamiliar to scientists. Whilst the next section will be dedicated to the analysis and visualization of large data sets, it is also important to consider how this data is stored and accessed.

Databases offer an attractive solution to both the storage of and the structured querying of data but it should be noted that there are many flavours of database architecture, and choosing the right one for your needs is crucial.

In general, databases are split into two camps: the relational databases based upon the Structured Query Language (SQL), and the non-relational databases (NoSQL). Each offers distinct pros and cons which should be weighed up when deciding on how to proceed. SQL databases offer complete transactional integrity, but lack flexibility. On the other hand, NoSQL databases offer a much more flexible structure, but do not guarantee transactional integrity. Additionally, its object orientated nature makes NoSQL databases more intuitive for storing different types of data which is all linked to one parent. A basic schema that shows how the flexible NoSQL format allows the storage of different types of data is shown in **Figure 5**. Due to its flexible framework and data types, the same underlying model has been successfully applied to projects of varying size and with varying needs and so, within our group, we favour the use of a NoSQL database (MongoDB) for the storage of the vast majority of our information.

The main characteristics of SQL and NoSQL databases are summarized in **Figure 6**, but it should be noted that a well implemented NoSQL database will outperform a poorly implemented SQL database, and vice versa.

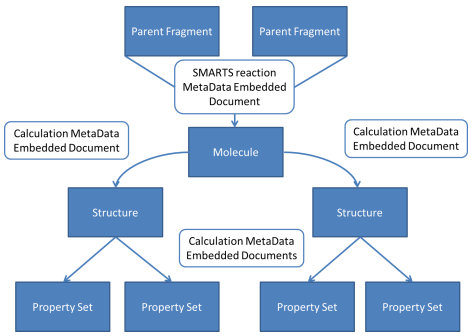

Figure 5

A basic schematic showing how different types of data (molecular descriptors, geometries, properties, calculation meta-data) can co-exist easily within a NoSQL format. Arrows represent bi-directional links between documents, and are embedded within the parent and child document. Additional meta-data is embedded within documents, which improves performance by reducing the number of steps required in a query.

One major benefit of interacting with data in a database format, whatever the flavor, is

that it affords the scientist access to a wide range of tools specifically designed for analyzing large amounts of data. All of the database architectures have their own shell for querying, and also allow access through a variety of popular scripting languages; large scale data analysis has never been easier.

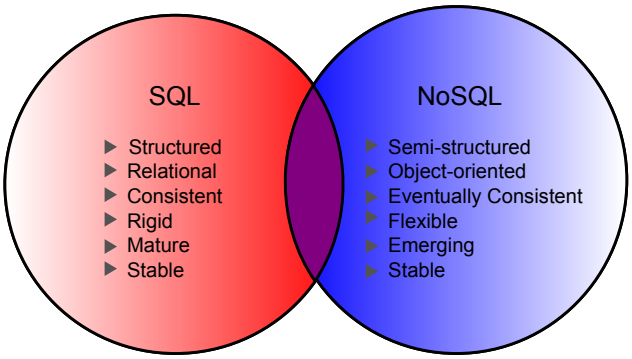

Figure 6

A comparison between SQL and NoSQL architectures; whilst SQL does allow transactional integrity, we believe that the enhanced flexibility of NoSQL databases make them the ideal choice for high throughput virtual screening since they can be easily modified to adapt to changing data and requirements.

Another advantage of NoSQL databases is the fact that their performance scales both horizontally and vertically. Whilst SQL databases need, on the whole, to be stored on one machine; NoSQL databases were designed to be split over many machines - a process known as sharding. The process of sharding spreads data across shards (servers), and since NoSQL architectures implement this natively, most will support balancing and query loading - resulting in good database performance with minimal maintenance costs. Since performance can be gained by simply adding more machines to the system, it represents a much cheaper option than having to purchase increasingly powerful machines as your database grows.

6. On the Analysis of Large Amounts of Chemical Data

It is highly desirable to analyze chemical data generated from high-throughput screening in a high throughput manner. The primary targets using high throughput data analysis is for (a) the suggestion, prioritization, and identification of top candidates for targeted synthesis and (b) uncovering sophisticated knowledge described as quantitative structure-property (activity) relationships (QSAR/QSPR) toward rationalization of selected candidates (3) and (65). More importantly, using high throughput data analysis creates unprecedented insight for future experiments. Both the generation of large-scale data and fast, yet accurate, data analysis in an unbiased manner are equally important in high throughput screening (66, 67).

It is recognized that due to the sheer size of the data produced in a high-throughput screen, using traditional data analysis techniques will impose a significant bottleneck upon

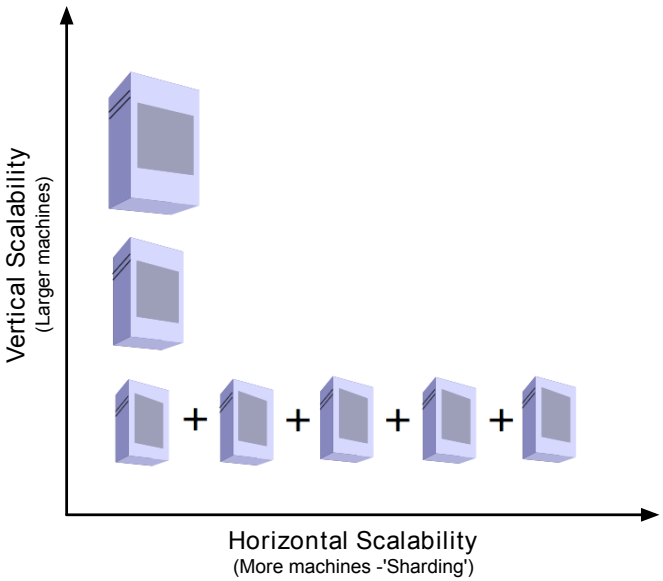

Figure 7

A diagramatic representation of vertical and horizontal scaling in database architectures. Since there is a non-linear scaling between computer power and cost, a vertical solution will be more expensive than its horizontal counterpart.

the screening procedure. This will in turn impose restrictions on the agility of the screening cycle to adapt to emerging trends and results. (68). We face exponentially growing requirements around organizing, summarizing, and interpreting such large amounts of high throughput data with respect to the vast chemical space to explore. In particular, there are three challenges in high throughput data analysis. First, the data we collect is increasingly voluminous, high-dimensional, complex, noisy, and diverse. Second, the chemical space to explore is essentially infinite and even a well-designed chemical library cannot cover all possible chemistry (69).

Currently, chemical data being generated through high throughput screening overwhelms traditional data analysis, undermining our ability to perform interactive and exploratory analysis and visualization of these data in post-processing. This situation will be exacerbated since it will grow by orders of magnitude in the near future due to the ever increasing hardware capability of computational resources. While high throughput screening still needs on some level to provide the role of traditional data analysis, which scrutinizes each individual data point, a further complication is the removal of bottlenecks to suggesting candidates and extracting useful information such as QSAR/QSPR by treating the data as a whole. To solve the issues addressed here, there are four main components to consider and perform in high throughput screening data analysis: processing, mapping and visualization, interpretation, and modeling.

Data processing: high throughput data analysis starts with ensuring the quality of data we collected for further discovery procedures. Multiple tasks such as cleaning, organization, normalization, and outlier detection need to be performed before a full data

exploration. When summarizing data, statistics including analyses of systematic error, correlation, or associations must be combined with simple visualizations such as heat maps (70). Before performing additional data analysis it is important to consider molecular descriptors that are encoded representation of molecules and useful for construction of QSAR/QSPR models (67) and (71). Unlike high throughput data analysis for inorganic materials, analysis of chemical data has distinct features in the development of descriptors. With the aid of graph theory, for instance, fingerprints and fragments have become the preferred descriptors for effective exploration of chemical space (3) and (53).

Data mapping and visualization: Mapping and visualization is one of the crucial components in high throughput data analysis because it is a direct way to depict chemical data and/or mined results with respect to chemical space to get insight into QSAR/QSPRs (72) and (73). However, simultaneous mapping of information is not always simple due to the nature of multi-dimensionality stemming from enumerated molecules, rules for bonds, functional groups, their chemical properties, and so forth. According to the similar property principle (74) and (75), similar chemical structures should have similar physicochemical properties. In that sense, it is crucial to choose proper measure of distance in a high-dimensional space to logically place molecules in property spaces defined by proper coordinates in high-dimensional visualization for visual mining (73), (76), (77), and (78). There are two broad categories of visualization for chemical information: direct visualization without data treatment and with data mined results. Among the former are plot matrix, parallel coordinates, heat maps, and other approaches (72). The typical approach for the latter includes dimensionality reduction to lower high-dimensions to manageable levels. This can be achieved by linear and non-linear dimensionality reduction, using, for example principal component analysis and diffusion map embedding. (79).

Data interpretation: Given the large quantity of data generated within a vast chemical space during high throughput screening, it is a natural choice to use data mining for identifying hot spots (e.g. interesting regions) where we are likely to find more candidates with desired functionalities from a vast chemical space (3) and (67). Data mining provides a flexible computational path for meeting the need to explore large amounts of chemical information. We often use various methods including classification, clustering, and prediction.

Data modeling for calibration: It has been previously mentioned that despite our best methods, experimental values are often not reproduced by computational techniques due to the inclusion of some systematic error. One way to correct for this is to calibrate the calculated results to experimental values. This results in data which is more intuitively analyzed by experimental collaborators.

The challenge in conducting high throughput screening is ensuring the pace of exploration of search space while keeping high levels of accuracy of calibration models (80). Given the two critical factors of speed and accuracy, high throughput screening mainly utilizes two types of approaches in advanced calibration modeling: hard modeling and soft modeling. To understand material's behaviors in a unified way, hard modeling captures different length scales of materials behaviors with chemistry- and physics-based theories and integrates such information (81). Although it provides highly accurate results, the modeling generally requires high computational cost. Exemplary approaches of it include ab-initio calculations and thermodynamic modeling.

In organic chemistry community, soft modeling has been a powerful approach not only for enhancing the accuracy of hard modeling (82) and (83) but also for making fast

and accurate property predictions (84). Soft modeling is based on statistical learning methods to seek heuristic relationships between data (81) and it often uses developed descriptors as well as knowledge extracted through the previous tasks to construct a cheap yet robust model enabling the establishment and deeper understanding of QSAR/QSPR (65). It includes regressions, artificial neural networks, multivariate analysis, and other machine learning algorithms (85). Unlike hard modeling, the predictive soft modeling is particularly valuable when physical/chemical models are not available. The models from soft modeling are relatively cheaper to construct than those fully-generated from expensive quantum chemical calculations, allowing accelerated screening procedures by replacing huge number of such theoretical investigations with heuristically developed QSAR/QSPRs.

A good example of predictive calibration approach is the application of Gibbs energy relationships to correlate electrode potentials of quinone/hydroquinone couples (41). Inspired by Dewar and Trinjajstic's early work (86), Huskinson *et al.* computed the differences in gas phase energy between oxidized and reduced states of quinone couples and successfully correlated those with measured redox potential values (41).

Note that robust calibration models should be developed from unskewed subspaces and be generalized to appropriately cover other possible chemistries. In other words, to ensure a higher prediction power for a wide-range of interesting candidates, the calibrated data needs to be distributed in a proper range to cover a larger chemical space of molecules to explore. It also should be noted that the quality of the training set determines the accuracy of calibration model when screening large number of new hypothetical molecules with limited or no experimental data in particular.

High throughput analysis of chemical data is a critical field of study in chemistry. It has emerged to address the issues associated with larger, more complex data sets. With increased understand of relationships between structure and property through high throughput data analysis, the most concrete outcome from data-driven models is to direct future experiments for high performance (e.g. output of desired properties) materials. More abstractly, application of such data-driven models will greatly enhance our understanding of basic physical and chemical principles (87). To those ends, data fusion and informatics platforms are importance pieces that take greater advantage of knowledge. Data fusion is an indispensable procedure in cutting-edge high throughput data analysis to link structures and properties (88). It can be more effectively completed when performing high throughput data analysis within the proper informatics platforms (68).

7. Future Directions

With the constant onslaught of new and improved computational hardware, and the increased uptake of deployment techniques such as distributed computing, we strongly believe that high throughput virtual screening has an important role to play in the future of materials science.

A key factor in this will be the resurgence of cheap, approximate techniques such as semi-empirical quantum methods, and QSPR approaches. Whilst the lower accuracy of these methods has resulted in a downturn in their usage in isolation, they gain a new value when included in a computational funnel, since only relative values are important, and known errors can be tolerated.

We predict that there will also be a sharp increase in the use of machine-learning techniques to act as a fast approximation for materials properties. Their ability to exploit

complex relationships between seemingly unconnected descriptors, and the fact that they can be easily trained against a small subset of chemical space that is directly relevant to the specific problem at hand makes them ideal for attacking these problems. Additionally, Bayesian methods have a knowledge not only of the result of the model, but a confidence in that answer. This can in turn be exploited to train the model on the fly, significantly increasing its value.

For these models to work, it is essential that there are good quality experimental results to calibrate against. It is therefore essential that experiments are performed in both an exploitative and an exploratory manner. That is to say, experiments which increase the knowledge of the local chemical space are as important as those which are focused on optimizing properties. This can only happen with continued implementation of the ideas of the Materials Genome Initiative, which aims to enable experimental and theoretical teams to work together to reap the benefits from the increased efficiency derived from working with a high throughput virtual screen.

With regards to experimental data, automated methods to collect the data, classify it with respect to experimental conditions and measurement type are required for further progress. Automated statistical methods to aid the assesment of reported vs. actual experimental error bars will allow for better calibration of theory to experiment.

Finally, the development of software tools aimed at the collaboration between theoretican and experimentalists while poring over the large datasets of hight-throughput virtual screens, and iteration of synthesis and computation will aid to bring screening methods to many more communities than the current group of scientists that practice it.

SUMMARY POINTS

1. A high throughput virtual screen is best defined by the philosophy employed in approaching the problem
2. Library generation is a compromise between including a big part of the molecular space and keeping the computational expenses tractable; a feedback loop between theory and experiment is recommended to keep the search in productive areas of chemical space
3. It is important to consider both the cost and the accuracy of calculations in a high throughput virtual screen; cheap methods have significant value in minimizing the number of molecules which are calculated using high-level methods
4. Identifying trends in the data is as important as identifying specific results
5. Calibrating results to experimental data can overcome deficiencies in specific methods
6. Exploratory, as well as exploitative, experimental results are crucial in the long term success of high throughput virtual screening

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be viewed as affecting the objectivity of this review

ACKNOWLEDGMENTS

The authors wish to thank Martin Blood-Forsythe and Suleyman Er for helpful discussions. The PI also wishes to acknowledge Samsung Electronics Co., Ltd. (SAIT); Dept. of Energy - through Grant DE-SC0008733; and Advanced Research Projects Agency (ARPA-E) through Grant Energy DE-AR0000348.

LITERATURE CITED

1. Reymond JL, van Deursen R, Blum LC, Ruddigkeit L. 2010. Chemical space as a source for new drugs. *MedChemComm* 1:30
2. Cedar G, Persson K. 2013. How Supercomputers Will Yield a Golden Age of Materials Science. *Scientific American* 309
3. Lipinski C, Hopkins A. 2004. Navigating chemical space for biology and medicine. *Nature* 432:855–861
4. Wermuth C. 2006. Selective optimization of side activities: the SOSA approach. *Drug Discov Today* 11:160–4
5. Wang M, Hu X, Beratan DN, Yang W. 2006. Designing Molecules by Optimizing Potentials. *J. Am. Chem. Soc.* 128:3228–3232
6. Balawender R, Welearegay MA, Lesiuk M, Proft FD, Geerlings P. 2013. Exploring Chemical Space with the Alchemical Derivatives. *J. Chem. Theory Comput.* 9:5327–5340
7. Tu M, Rai BK, Mathiowetz AM, Didiuk M, Pfefferkorn JA, et al. 2012. Exploring Aromatic Chemical Space with NEAT: Novel and Electronically Equivalent Aromatic Template. *Journal of Chemical Information and Modeling* 52:1114–1123
8. Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN. 2013. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* 135:7296–7303
9. Ehrlich HC, Henzler AM, Rarey M. 2013. Searching for Recursively Defined Generic Chemical Patterns in Nonenumerated Fragment Spaces. *Journal of Chemical Information and Modeling* 53:1676–1688
10. Hoksza D, Škoda P, Voršilák M, Svozil D. 2014. Molpher: a software framework for systematic chemical space exploration. *Journal of Cheminformatics* 6:7
11. Fink T, Bruggesser H, Reymond JL. 2005. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angew. Chem. Int. Ed.* 44:1504–1508
12. Blum LC, Reymond JL. 2009. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* 131:8732–8733
13. Ruddigkeit L, van Deursen R, Blum LC, Reymond JL. 2012. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* 52:2864–2875
14. Taniguchi M, Du H, Lindsey JS. 2011. Virtual Libraries of Tetrapyrrole Macrocycles. Combinatorics Isomers, Product Distributions, and Data Mining. *Journal of Chemical Information and Modeling* 51:2233–2247
15. Yu MJ. 2011. Natural Product-Like Virtual Libraries: Recursive Atom-Based Enumeration. *Journal of Chemical Information and Modeling* 51:541–557
16. Massarotti A, Brunco A, Sorba G, Tron GC. 2014. ZINClick: A Database of 16 Million Novel Patentable, and Readily Synthesizable 1,4-Disubstituted Triazoles. *Journal of Chemical Information and Modeling* 54:396–406
17. Koutsoukas A, Paricharak S, Galloway WRJD, Spring DR, IJzerman AP, et al. 2014. How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *Journal of Chemical Information and Modeling* 54:230–242
18. Roth HJ. 2005. There is no such thing as 'diversity'! *Current Opinion in Chemical Biology*

9:293–295
19. Riniker S, Landrum GA. 2013. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics* 5:26
20. Maggiora G, Vogt M, Stumpfe D, Bajorath J. 2014. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* 57:3186–3204
21. Gillet V, Johnson A, Mata P, Sike S, Williams P. 1993. SPROUT: a program for structure generation. *J Comput Aided Mol Des* 7:127–53
22. Pearlman D, Murcko M. 1996. CONCERTS: dynamic connection of fragments as an approach to de novo ligand design. *J Med Chem* 39:1651–63
23. Schneider G, Lee M, Stahl M, Schneider P. 2000. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J Comput Aided Mol Des* 14:487–94
24. Gillet V, Willett P, Fleming P, Green D. 2002. Designing focused libraries using MoSELECT. *J Mol Graph Model* 20:491–8
25. Vinkers H, de JM, Daeyaert F, Heeres J, Koymans L, et al. 2003. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J Med Chem* 46:2765–73
26. Brown N, McKay B, Gasteiger J. 2004. The de novo design of median molecules within a property range of interest. *Journal of Computer-Aided Molecular Design* 18:761–771
27. Nicolaou C, Brown N, Pattichis C. 2007. Molecular optimization using computational multiobjective methods. *Curr Opin Drug Discov Devel* 10:316–24
28. Liu Q, Masek B, Smith K, Smith J. 2007. Tagged Fragment Method for Evolutionary Structure-Based De Novo Lead Generation and Optimization. *Journal of Medicinal Chemistry* 50:5392–5402
29. Dey F, Caflisch A. 2008. Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization. *Journal of Chemical Information and Modeling* 48:679–690
30. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, et al. 2012. Automated design of ligands to polypharmacological profiles. *Nature* 492:215–220
31. Osedach TP, Andrew TL, Bulović V. 2013. Effect of synthetic accessibility on the commercial viability of organic photovoltaics. *Energy Environ. Sci.* 6:711
32. O'Boyle NM, Campbell CM, Hutchison GR. 2011. Computational Design and Selection of Optimal Organic Photovoltaic Materials. *J. Phys. Chem. C* 115:16200–16210
33. Kanal IY, Owens SG, Bechtel JS, Hutchison GR. 2013. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* 4:1613–1623
34. Bertz SH. 1981. The first general index of molecular complexity. *J. Am. Chem. Soc.* 103:3599–3601
35. Boda K, Johnson A. 2006. Molecular complexity analysis of de novo designed ligands. *J Med Chem* 49:5869–79
36. Bonnet P. 2012. Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists. *European Journal of Medicinal Chemistry* 54:679–689
37. Podolyan Y, Walters MA, Karypis G. 2010. Assessing Synthetic Accessibility of Chemical Compounds Using Machine Learning Methods. *Journal of Chemical Information and Modeling* 50:979–991
38. Warr WA. 2014. A Short Review of Chemical Reaction Database Systems Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inf.* 33:469–476
39. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sanchez-Carrera RS, et al. 2011. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* 2:2241–2251
40. Olivares-Amaya R, Amador-Bedolla C, Hachmann J, Atahan-Evrenk S, Sanchez-Carrera RS, et al. 2011. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* 4:4849

41. Huskinson B, Marshak MP, Suh C, Er S, Gerhardt MR, et al. 2014. A metal-free organic–inorganic aqueous flow battery. *Nature* 505:195–198
42. Goushi K, Yoshida K, Sato K, Adachi C. 2012. Organic light-emitting diodes employing efficient reverse intersystem crossing for triplet-to-singlet state conversion. *Nature Photonics* 6:253–258
43. Zhang Q, Li B, Huang S, Nomura H, Tanaka H, Adachi C. 2014. Efficient blue organic light-emitting diodes employing thermally activated delayed fluorescence. *Nature Photonics* 8:326–332
44. Korth M. 2014. Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: evaluation of electronic structure theory methods. *Phys. Chem. Chem. Phys.* 16:7919
45. https://www.materialsproject.org/ - The Materials Project. *accessed 2014-09-10*
46. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, et al. 2013. Python Materials Genomics (pymatgen): A robust open-source python library for materials analysis. *Computational Materials Science* 68:314–319
47. Kresse G, Furthmüller J. 1996. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science* 6:15–50
48. Blöchl PE. 1994. Projector augmented-wave method. *Physical Review B* 50:17953–17979
49. Perdew JP, Burke K, Ernzerhof M. 1996. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* 77:3865–3868
50. Anisimov VI, Zaanen J, Andersen OK. 1991. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Physical Review B* 44:943–954
51. Jain A, Hautier G, Moore CJ, Ong SP, Fischer CC, et al. 2011. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science* 50:2295–2310
52. Jain A, Ong SP, Hautier G, Chen W, Richards WD, et al. 2013. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* 1:011002
53. Hachmann J, Olivares-Amaya R, Jinich A, Appleton AL, Blood-Forsythe MA, et al. 2014. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project. *Energy Environ. Sci.* 7:698
54. World Community Grid - https://secure.worldcommunitygrid.org/index.jsp. *accessed 2014-09-10*
55. Scharber MC, Mühlbacher D, Koppe M, Denk P, Waldauf C, et al. 2006. Design Rules for Donors in Bulk-Heterojunction Solar Cells—Towards 10 % Energy-Conversion Efficiency. *Adv. Mater.* 18:789–794
56. Shockley W, Queisser HJ. 1961. Detailed Balance Limit of Efficiency of p-n Junction Solar Cells. *J. Appl. Phys.* 32:510
57. The Harvard Clean Energy Project Database - https://cepdb.molecularspace.org/. *accessed 2014-09-10*
58. Kolossváry I, Guida WC. 1996. Low Mode Search. An Efficient Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides. *J. Am. Chem. Soc.* 118:5011–5019
59. Sadowski J, Gasteiger J, Klebe G. 1994. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *Journal of Chemical Information and Modeling* 34:1000–1008
60. Schrodinger LLC New York N. 2014. Macromodel
61. Mayo SL, Olafson BD, Goddard WA. 1990. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* 94:8897–8909
62. Folding at home - http://folding.stanford.edu. *accessed 2014-09-01*
63. SETI at home - http://setiathome.berkeley.edu/. *accessed 2014-09-01*
64. CamGrid - http://www.ucs.cam.ac.uk/scientific/camgrid. *accessed 2014-09-10*
65. Parker CN, Shamu CE, Kraybill B, Austin CP, Bajorath J. 2006. Measure mine, model, and

manipulate: the future for HTS and chemoinformatics? *Drug Discovery Today* 11:863–865
66. Tamura SY, Bacha PA, Gruver HS, Nutt RF. 2002. Data Analysis of High-Throughput Screening Results: Application of Multidomain Clustering to the NCI Anti-HIV Data Set. *Journal of Medicinal Chemistry* 45:3082–3093
67. Harper G, Pickett SD. 2006. Methods for mining HTS data. *Drug Discovery Today* 11:694–699
68. Ling X. 2008. High Throughput Screening Informatics. *Combinatorial Chemistry & High Throughput Screening* 11:249–257
69. Medina-Franco J, Martinez-Mayorga K, Giulianotti M, Houghten R, Pinilla C. 2008. Visualization of the Chemical Space in Drug Discovery. *CAD* 4:322–333
70. Asli N. Goktug SCC, Che T. 2013. In *Drug Discovery*. InTech
71. García-Domenech R, Gálvez J, de Julián-Ortiz JV, Pogliani L. 2008. Some New Trends in Chemical Graph Theory. *Chem. Rev.* 108:1127–1169
72. Suh C, Sieg SC, Heying MJ, Oliver JH, Maier WF, Rajan K. 2009. Visualization of High-Dimensional Combinatorial Catalysis Data. *J. Comb. Chem.* 11:385–392
73. Awale M, van Deursen R, Reymond JL. 2013. MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank ChEMBL, PubChem, GDB-11, and GDB-13. *Journal of Chemical Information and Modeling* 53:509–518
74. Klopmand G. 1992. Concepts and applications of molecular similarity by Mark A. Johnson and Gerald M. Maggiora, eds., John Wiley & Sons, New York, 1990, 393 pp. Price: $65.00. *J. Comput. Chem.* 13:539–540
75. Willett P, Barnard J, Downs G. 1998. Chemical Similarity Searching. *Journal of Chemical Information and Modeling* 38:983–996
76. Chen X, Reynolds C. 2002. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *Journal of Chemical Information and Modeling* 42:1407–1414
77. Godden JW, Bajorath J. 2006. A Distance Function for Retrieval of Active Molecules from Complex Chemical Space Representations. *Journal of Chemical Information and Modeling* 46:1094–1097
78. Haranczyk M, Holliday J. 2008. Comparison of Similarity Coefficients for Clustering and Compound Selection. *Journal of Chemical Information and Modeling* 48:498–508
79. Coifman RR, Lafon S. 2006. Diffusion maps. *Applied and Computational Harmonic Analysis* 21:5–30
80. Platts J, Butina D, Abraham M, Hersey A. 1999. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *Journal of Chemical Information and Modeling* 39:835–845
81. Liu ZK, Chen LQ, Rajan K. 2006. Linking length scales via materials informatics. *JOM* 58:42–50
82. Balabin RM, Lomakina EI. 2011. Support vector machine regression —an alternative to artificial neural networks for the analysis of quantum chemistry data? *Phys. Chem. Chem. Phys.* 13:11710
83. Balabin RM, Lomakina EI. 2009. Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. *J. Chem. Phys.* 131:074104
84. Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. 2013. Accelerating materials property predictions using machine learning. *Scientific Reports* 3
85. Rajan K, Suh C, Mendez PF. 2009. Principal component analysis and dimensional analysis as materials informatics tools to reduce dimensionality in materials science and engineering. *Statistical Analysis Data Mining* 1:361–371
86. Dewar M, Trinajstic N. 1969. Ground states of conjugated molecules—XIV. *Tetrahedron* 25:4529–4534
87. Bajorath J. 2001. Selected Concepts and Investigations in Compound Classification Molecular Descriptor Analysis, and Virtual Screening. *Journal of Chemical Information and Modeling*

41:233–245
88. Searls DB. 2005. Data integration: challenges for drug discovery. *Nat Rev Drug Discov* 4:45–58