Journal of Medicinal Chemistry

Perspective

Cite This: J. Med. Chem. 2019, 62, 1116−1124

pubs.acs.org/jmc

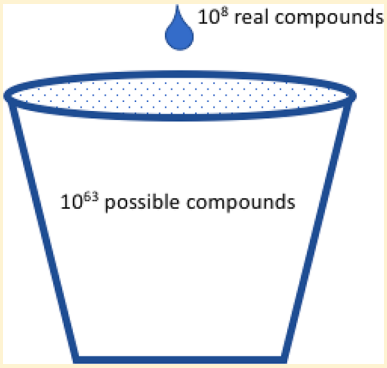
Virtual Chemical Libraries

Miniperspective

W. Patrick Walters*

Relay Therapeutics, 215 First Street, Cambridge, Massachusetts 02142, United States

ABSTRACT: Advances in computer processing speed and storage capacity have enabled researchers to generate virtual chemical libraries containing billions of molecules. While these numbers appear large, they are only a small fraction of the number of organic molecules that could potentially be synthesized. This review provides an overview of recent advances in the generation and use of virtual chemical libraries in medicinal chemistry. We also consider the practical implications of these libraries in drug discovery programs and highlight a number of current and future challenges.



INTRODUCTION

Many have remarked that medicinal chemistry is entering the age of "big data".[1−3] In 2015, Chemical Abstracts Service reported the registration of its 100 millionth compound.[4] Medicinal chemists now have access to publicly available and proprietary databases with chemical structures and biological data for millions of compounds. Table 1 lists a few of the more popular databases. While some of these databases appear large, their collective size is still miniscule compared to the universe of potential drug-like compounds. One approach to bridging the gap between existing compounds and compounds that could be synthesized is to computationally generate libraries of virtual compounds and evaluate these libraries using a variety of virtual screening methods. In this review, we will examine some recent developments in the generation and usage of virtual chemical libraries.

We will begin by examining how different researchers have estimated the size of chemical space. As the basis for further discussions of virtual chemical libraries, it is instructive to review the methods that have been used to estimate the number of possible organic compounds. We will follow this with an overview of methods designed to fully enumerate all possible small organic compounds. While it is useful to be able to examine the entirety of chemical space, many would ultimately prefer to focus on molecules that can be synthesized. In recent years, a number of groups have constructed large virtual libraries based on precedented chemical reactions, often favoring chemistries that are suitable for automation. These efforts have led to a number of proprietary and publicly available databases. While databases of synthetically accessible compounds are useful, many groups lack the laboratory resources necessary to carry out even the most basic organic syntheses. Fortunately, commercial vendors have begun to offer databases of readily synthesizable compounds that can be delivered at a relatively low cost in a matter of a few weeks. An alternative approach to virtual chemical libraries has been motivated by recent advances in artificial intelligence. We will provide an overview of methods known as generative models that provide a means of performing a directed search through large chemical spaces. Finally, we will conclude with a discussion of the practical aspects of using large virtual libraries as well as some prospects for future developments.

HOW LARGE IS CHEMICAL SPACE?

As mentioned above, the number of organic compounds that could be synthesized is enormous, with numbers between 10^{30} and 10^{60} being routinely cited. Since these estimates vary widely, one may wonder how these numbers were derived. In this section, we will provide some background and explain how a number of authors arrived at their estimates of the size of chemical space.

One of the earliest estimates of the size of chemical space came from Weininger[5] who used a set of 150 substituents on hexane to estimate the size of small molecule organic chemical space at 10^{29}. Independently, Bohacek and co-workers[6] estimated the number of possible organic molecules by imagining the buildup of a linear chain consisting of C, N, O, or S atoms. By considering the possibility of multiple bonds, the authors assumed an average of 6 possible options for each atom added to the growing chain. Considering a maximum of 30 atoms led to a total of 6^{30} or 2×10^{23} molecules. The addition of ring closures led to an estimate of 10^{40} molecules containing up to 4 rings and 10 branch points. The combination of linear and branched pieces led to a combined estimate of 10^{63} molecules.

Received: July 2, 2018
Published: August 27, 2018

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124

Table 1. Large, Publicly Available Chemical Libraries

database	purpose	number of compounds	reference	URL
ChEMBL	compounds and data from medicinal chemistry literature	2 million	43, 74	https://www.ebi.ac.uk/chembl/
PubChem	compounds and data from academic screening centers	90 million	75, 76	https://pubchem.ncbi.nlm.nih.gov/
ChemSpider	collection of chemical structures from multiple data sources	63 million	77, 78	http://www.chemspider.com/
ZINC	commercially available compounds for virtual screening	980 million	79	http://zinc15.docking.org/
SureChEMBL	named examples from chemical patents	17 million	80	https://www.surechembl.org/search/

In a subsequent paper, Ertl[7] used an analysis of organic substituents to estimate the number of possible organic molecules at somewhere between 10^{20} and 10^{24}. In this analysis, the author began with a set of 3 million commercially available compounds and used an algorithm to extract 849 574 unique substituents with 12 or fewer heavy atoms. This analysis showed that only 50 substituents occurred in at least 1% of the molecules and 438 substituents were present in 0.1% of molecules. These results were consistent with an earlier analysis of drugs and clinical candidates.[8] Ertl found a linear relationship between the log of the number of molecules processed and the log of the number of substituents identified. He then estimated the total number of substituents as 3.1 million. With the number of possible substituents in hand, he created a simple model to determine the number of molecules that can be constructed by adding 2 substituents to a molecular scaffold. Using this analysis, he determined that 5.2×10^{19} molecules can be formed from building blocks with up to 12 atoms. By use of a formula that allowed 3 substituents, the number of possible molecules grew to 6.7×10^{23}.

In a 2007 paper, Ogata and co-workers[9] used combinations of protein fragments to generate drug-like molecules and subsequently estimate the size of lead-like chemical space. The authors began by extracting the coordinates of protein atoms from the Protein Data Bank (PDB). These structures were then decomposed into fragments, which were classified as rings, linkers, and side chains. Fragments were then combined using a set of rules that ensured consistent geometries and avoided violations of valence constraints. Results were then filtered for lead-likeness using criteria established by Lipinski et al.[10] In addition, compounds containing potentially objectionable functionality (toxic, reactive, etc.) were removed using rules established by Rishton.[11] On the basis of this analysis, the authors posit that there are more than 10^8 lead-like compounds composed of C, O, N, S, and Cl atoms.

In a more recent data driven approach, Polishchuk and co-workers[12] used the distribution of compounds in the generated database (GDB), which enumerates all possible organic molecules with up to 17 atoms (see below), to estimate the size of chemical space. By plotting the number of heavy atoms against the log of the number of enumerated molecules, the authors derived this equation,

$$\log M = 0.584N \log(N) + 0.356$$

where M is the number of molecules and N is the number of atoms in a molecule. By using the equation above to extrapolate to 36 atoms, the authors arrived at an estimate of 10^{33} drug-like molecules.

On the basis of the work described above, it is apparent that there is no "correct" number of possible molecules. However, based on several different approaches, it is clear that the number of drug-like compounds synthesized to date is a tiny fraction of the number of possible molecules. There are still vast areas of chemistry space to explore, and one approach to exploring that space is to build and evaluate large virtual libraries. In the next sections we will review a number of approaches to constructing virtual libraries that enable this exploration.

■ ENUMERATING CHEMICAL SPACE

One approach to the identification of novel molecules is to enumerate all possible molecules that obey the rules of valence. This is the objective of the Chemical Space Project,[13] which has been developed by the Reymond group at the University of Berne. This group has developed algorithms for exhaustively enumerating molecular skeletons. In the first part of this process, molecules are enumerated and those with excessive ring strain are eliminated. The molecular skeletons are then refined by enumerating bond orders and atom types including C, N, O, S, and halogens. Molecules that violate rules of valence or contain potentially unstable bonds are then eliminated yielding a set of molecules that has been made publicly available in a database known as the GDB. The first version of this database, GDB-11,[14] contained 26.4 million molecules with up to 11 heavy atoms. A subsequent version, GDB-13,[15] released 2 years later raised the limit to 13 heavy atoms and encompassed 970 million molecules. The most recent release, GDB-17,[16] contains molecules with up to 17 heavy atoms and has ballooned to 166 billion molecules. Portions of the GDB database have been used for virtual screening and have provided hits for a number of drug discovery targets including the NMDA receptor[17,18] and nicotinic acetylcholine receptor (nAChR).[19,20]

Rather than fully enumerating all possible compounds, others have created sublibraries that are designed to be representative of the larger chemical space. One such effort is the work of Virshup and co-workers,[21] who devised a procedure known as Algorithm for Chemical Space Exploration with Stochastic Search (ACSESS) to explore untapped regions of chemical space. The ACSESS method begins with a small set of simple molecules and modifies these molecules using crossover and mutation operations similar to those found in genetics. In the crossover operation, molecules are first fragmented by removing acyclic bonds. The resulting fragments from different molecules are then recombined to create new molecules. Mutations introduce random changes to molecules by adding or removing atoms, introducing ring closures, or changing bond orders or atom types (e.g., C to N). The resulting molecules are then modified in a similar fashion, and the process is repeated until a predefined number of iterations has been reached. During this process, molecular descriptor calculations are used to "steer" the evolution of the library into or away from particular regions of chemical space. Once library generation was complete, the authors used principal component analysis (PCA) and a self-organizing map (SOM) to compare the chemical space covered by an 8.9 million compound library, generated by ACSESS, with a number of databases containing medicinally relevant com-

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124

pounds. The authors found that the 30 million compounds in the PubChem database covered only 2% of the chemical space covered by the database constructed using ACSESS. A comparison of a maximally diverse set of 10 000 compounds enumerated with ACSESS with the 970 million compounds in GDB-13 showed that the ACSESS library was able to cover a similar chemical space with 10^4 fewer compounds.

In a related approach, called "Chemical Space Travel" (CST),[22] the Reymond group has developed an algorithm that creates a "path" in chemical space between two molecules. The program begins with a source molecule and generates more than 500 000 molecules on a trajectory to a target molecule. CST begins by making random mutations, similar to those described above for ACSESS, to a set of starting chemical structures. These mutations include atom type exchange, atom additions and deletions, bond saturation, bond rearrangement, and aromatic ring addition. The process is repeated, and the "mutant" molecules that are most similar to the target molecule are mutated in a similar fashion. This process is repeated until the resulting mutant molecules are sufficiently similar to the target molecule. In an example, the authors demonstrate how CST can be used to create molecules that are hybrids of existing AMPA receptor ligands.

VIRTUAL "ON-DEMAND" LIBRARIES

A number of recent papers have highlighted the fact that the majority of compounds synthesized in pharmaceutical companies are based on a relatively small number of chemical reactions.[23−25] This fact can, of course, be considered a double-edged sword. One can argue that reliance on a small number of reactions reduces the overall diversity of the resulting molecules. On the other hand, these widely used reactions are reliable and have led to the availability of large numbers of applicable commercial building blocks. The emergence of robust reactions and numerous diverse reagents has led a number of groups to create large virtual libraries based on validated chemical reactions. In many cases, these computational efforts have been motivated by large investments in systems for performing automated synthesis.

Scientists at Pfizer used data from 12 years of parallel synthesis to construct a large virtual library known as the Pfizer Global Virtual Library (PGVL).[26,27] A set of more than 1000 reactions used to synthesize more than 2 000 000 compounds for the Pfizer compound collection were used to define specific rules for reagent selection. Reagent structures were then transformed into a format suitable for library enumeration. In order to facilitate synthesis, the enumeration system also contains a sophisticated set of rules to resolve ambiguities where multiple reacting centers are present. An enumeration of the full set of 1244 reactions with reagents from the Pfizer compound collection led to 10^{14} compounds. A further expansion of the same set of reactions with a larger set of commercial reagents led to a library with 10^{18} compounds.

In a somewhat similar fashion, a group at Eli Lilly has developed the Proximal Lilly Collection (PLC),[28] a set of molecules that can be rapidly synthesized using Lilly's robotic synthesis tools and readily available starting materials. The system utilizes a set of 10 annotated reaction types to generate custom virtual libraries consisting of approximately 10 million molecules. These libraries are then subjected to a variety of virtual screening procedures. Larger virtual libraries on the order of 1 billion molecules can also be generated to meet the needs of specific drug discovery projects. Searches in these large virtual libraries will identify specific molecules and their associated synthetic routes. A parallelized similarity search can also be used to identify analogs for SAR expansion. As a validation study, the Lilly group searched the PLC for marketed drugs (3.2% found), molecules from the Lilly collection (23% found), and PubChem (20% found).

Academic groups have also engaged in the development of reaction-oriented virtual libraries. Chevillard and co-workers have developed SCUBIDOO,[29] a freely available database of molecules that could be generated using tractable synthetic reactions. The authors generated a library of 21 million products by carrying out virtual reactions on a pool of approximately 8000 commercially available building blocks using a set of 58 robust reactions originally published by Hartenfeller.[30] In order to facilitate synthesis, the authors have provided synthesis instructions, along with descriptions of potential side reactions. This documentation should streamline practical application of the technology. Realizing that some users may not have the computational capabilities to process a large virtual library, the authors have used sampling approaches to provide small, medium, and large (1000, 100K, 1M) subsets of the library.

The results of a similar effort were recently published by Humbeck and co-workers who created CH*I*PMUNK,[31] a library of small molecules with property profiles appropriate for modulators of protein−protein interactions. The CH*I*PMUNK library consists of 95 million compounds that were derived from predefined virtual reactions carried out using sets of commercially available building blocks. The authors report the number of virtual compounds generated using three different sets of reactions and two different sets of reagents. The first set of reagents (commercial reagents) is derived from the eMolecules[32] and MolPort[33] collections, and the second, ZINC reagents, is derived from the publicly available ZINC database. The numbers of virtual products generated are shown in Table 2.

Table 2. Number of Virtual Products Generated by Reaction Types Employed in the CH*I*PMUNK Library

reaction type	commercial reagents	ZINC reagents
heterocycle forming	10 million	11 million
medicinal chemistry	36 million	18 million
multicomponent	20 million	1 million

The majority of the virtual molecules in the CH*I*PMUNK library violate Lipinski's rule of 5,[10] primarily due to molecular weights greater than 500 g/mol. However, as the authors point out, molecules that modulate protein−protein interactions (PPI) tend to be larger and more lipophilic than typical drug molecules. On the basis of the similarity of property profiles to those found in known PPI inhibitors, the authors propose that the CH*I*PMUNK library would be most appropriate as a tool for virtual screening of PPIs. One novel aspect that increases the utility of the CH*I*PMUNK library is the inclusion of cluster identifiers, which enable rapid grouping of related compounds. Due to algorithmic limitations, it is very difficult to cluster chemical databases with more than 10 million compounds. The team behind the CH*I*PMUNK database utilized a novel method known as StruClus[34] to cluster this database of 95 million compounds. This clustering method was also used to compare the chemical space coverage of the CH*I*PMUNK database with the building block databases and the ChEMBL

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124

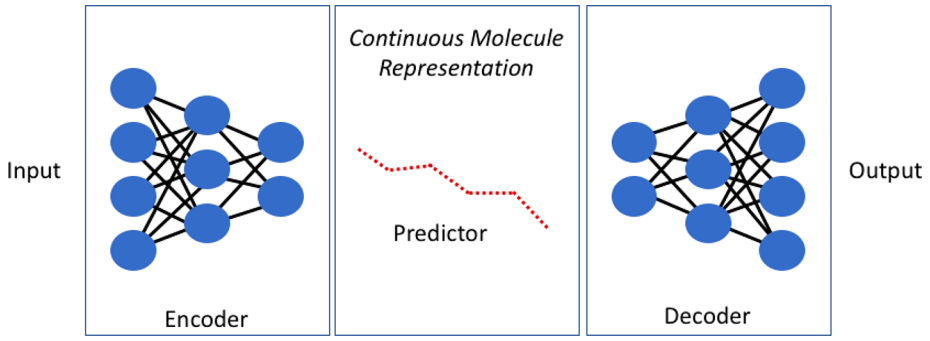


Figure 1. Architecture of a generative model.

database of medicinal chemistry compounds. The cluster analysis showed that the CHIPMUNK database covered different chemical space than either ChEMBL or the building block databases.

As mentioned in the Introduction, virtual libraries based around chemical reactions may only be useful to groups with the synthetic chemistry expertise and resources to synthesize the compounds. A recent trend, which may at least partially fill this gap, is the availability of "synthesis on demand" libraries from commercial vendors. A prime example of this concept is the REAL database from Enamine,[35] a screening compound supplier based in the Ukraine. The current version of the REAL (readily accessible) library consists of more than 680 million compounds based around chemical reactions that the company has developed and validated. According to the Enamine Web site, a set of 100 compounds can be delivered in 2−3 weeks. As will be discussed below, databases of this size are far beyond the magnitude of databases typically considered for virtual screening. New computational methods will have to be developed in order to fully take advantage of these multimillion compound databases.

GENERATIVE MODELS

For the past 30 years, virtual screening methods have fundamentally worked in one of two ways. In the first, a predefined list of molecules is presented to a computational method that evaluates the molecules in some context (e.g., molecular similarity or protein−ligand docking) and outputs a score. This score is then used to prioritize molecules for synthesis or purchase. In a second approach, commonly known as de novo design,[36] molecules are iteratively built up from smaller fragments. This building process, which can proceed one atom or one molecular fragment at a time, is directed by a scoring function similar to those used in virtual screening. Over the past few years, a new set of de novo design methods known as generative models[37] have appeared. These methods take advantage of recent advances in deep machine learning that have been applied in areas such as text analytics and image analysis.

A generative model typically consists of three components, as illustrated in Figure 1: an encoder that converts a set of molecules to a continuous vector representation, a decoder that converts the continuous vector representation back to a molecule, and a predictor that calculates one or more properties for vectors from the continuous representation. These three components are linked through a computational architecture known as a neural network. By using the predictor to calculate values in the vector representation, one can exploit gradients to direct exploration toward molecules with a desired profile. One can draw an analogy to widely publicized machine learning methods that are capable of generating images of human faces that, while appearing genuine, are actually artificial. In these methods, a neural network is initially presented with a set of images of human faces (often those of celebrities). The program then "learns" a continuous representation that captures the essence of the images. This continuous representation can then be decoded to generate a new set of images that, while appearing photorealistic, do not depict an actual person.

One of the first applications of generative models to molecular design was published in 2018 by Gómez-Bombarelli and co-workers.[38] In this paper, the authors used a text representation of molecules as SMILES strings to take advantage of previous neural network architectures that had been optimized to deal with English text. An autoencoder was used to transform SMILES strings from the ZINC database of commercially available screening compounds and from QM9,[39] a database of quantum chemical properties for molecules derived from the GDB-13 database (see above) to create a reduced dimensionality vector representation. A decoder was subsequently used to generate new molecules from the continuous representation. As a proof of concept, the authors coupled the neural network that generates molecules to a second network that calculated properties including the octanol−water partition coefficient (log P), synthetic accessibility score (SAS), and quantitative estimation of drug-likeness (QED). By coupling the two networks, the authors were able to sample regions of chemical space with a desired property profile.

A similar approach was applied by Segler and co-workers[40] who used a generative model to design molecules predicted to have antibacterial activity. In this work, the authors used a technique known as transfer learning to compensate for the relatively small number of antibacterial molecules available to train the model. In transfer learning, a large, more general data set is used to initially train a model, then a smaller more specific data set is used for fine-tuning. This technique has been widely applied in image analysis. For instance, a set of millions of general images (e.g., cat pictures) from the Internet may be used to initially train a model designed to identify tumors from radiology images. Once the model has learned critical elements of images (e.g., edge detection), the model can be fine-tuned on a set of radiology images. In a similar fashion, a generative model can be initially trained on large chemical databases and fine-tuned on specific sets of molecules. The generative model developed by Segler was initially trained on 1.4 million SMILES strings from the ChEMBL database. Model tuning was then carried out using

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124

data sets of chemical structures and associated antibacterial activity derived from the ChEMBL database. As a control experiment, the authors used a generative model to simulate the synthesis, design, test cycle commonly employed in drug discovery. An examination of the molecules produced by the generative models showed that the model was able to produce 14% and 28% of the molecules in held-out test sets of known antibacterial compounds.

It should be pointed out that generative models are a new and emerging area of research that is now being pursued by multiple groups, and a complete overview of the topic is beyond the scope of this review. The journal *Molecular Informatics*[37] recently devoted an entire issue to the topic. While a number of groups have been able to generate molecules that scored well in predictive models, there have only been a couple of examples of generative models being used to design molecules that have been shown to be active in biological assays. In one recent example, Putin and co-workers[41] trained a generative model on a filtered set of 5K kinase inhibitors. The structures produced by the generative model were then used as seeds for similarity searches which were used to select 50 commercially available compounds for purchase. Seven of the purchased molecules showed percent inhibition between 50% and 101% in a kinase panel assay.

In a recent paper, Merk and co-workers[42] have gone one step further and synthesized and tested a set of molecules produced by a generative model. These authors initially trained a generative model with a set of more than 500K biologically active molecules from the ChEMBL[43] database. As described above, the authors then used transfer learning to fine-tune their model with a set of 25 compounds having antagonistic activity against either the retinoid X receptors (RXR) or peroxisome proliferator-activated receptors (PPAR). The resulting molecules, which were built up from a carboxylic acid starting point, were then filtered through a target prediction model to identify potential RXR and PPAR agonists. On the basis of this analysis, five compounds were synthesized and tested for RXR and PPAR activity. Of the five compounds tested, four had EC_{50} values ranging between double digit nM and double digit μM.

Generative models are also bound to encounter one of the primary problems encountered with earlier de novo design programs. While these programs are capable of generating molecules that appear capable of binding to a particular site, the generated molecules are often very difficult to synthesize. One answer to the "synthesizability dilemma" has been to incorporate a synthetic accessibility score[44−47] into the process of exploring chemical space. Unfortunately, these synthesizability scores tend to employ rather crude heuristics that identify commonly formed bond types. The methods are often weakly validated and cannot be counted on for reliable estimates of synthetic accessibility. One interesting prospect, however, would be to couple a generative model with one of the very successful recent generation of deep neural networks designed to identify optimal synthetic routes.[48−50]

■ VISUALIZING CHEMICAL SPACE

One of the challenges associated with large virtual libraries is visualizing regions of chemical space occupied by a particular library. In many cases, we may want to compare the chemical space covered by one library with the space covered by another. An intuitive "map" would enable us to compare libraries as well as identify regions that contain chemically similar compounds. Once we have such a map, we could use color to indicate a number of other attributes such as biological activity or calculated properties.

When visualizing chemical space, molecules are typically represented as vectors. These vectors are often either sets of calculated properties[51−53] or molecular fingerprints[54−56] where a molecule is represented by a vector of 1s and 0s representing the presence or absence of specific features. One of the most common chemical space representations is to project these high dimensional vectors into two or three dimensions using a technique such as principal component analysis (PCA).[57,58] One early example of such an approach is the ChemGPS method published by Oprea. In this method, a set of 72 molecular descriptors were calculated and projected into a three-dimensional space. Each of the projected points represented one compound. In order to facilitate intuitive navigation in the space, a set of "satellite" molecular cores were identified. Another method commonly used for visualizing chemical space is the self-organizing map (SOM).[59,60] In an SOM, the positions of vectors representing molecules are iteratively adjusted so that points in a two-dimensional grid representing similar molecules are placed close together and points representing dissimilar molecules are far apart. While PCA and SOM are useful approaches for representing smaller compound sets (up to a million compounds), they do not scale to large virtual libraries with tens to hundreds of millions of compounds. In addition to the time required to calculate the maps, the sheer number of points on a plot can make interpretation difficult. Both of these methods also have the drawback that the projection from a high dimensional vector into a two or three-dimensional space that can be visualized may significantly distort the actual distances.

More recently, a number of groups have developed methods that are more appropriate for the visualization of virtual libraries containing millions of molecules. It is perhaps not surprising that some of this work originated with the Reymond group who created the 166 billion compound GDB-17 database. This group has generated a Web-based software tool known as FUn[61] that is optimized for the visualization of very large chemical libraries. An impressive demo on the FUn Web site shows an interactive 3D visualization of 17 million compounds from the SureChEMBL database. The visualization in the FUn program is facilitated by the use of a novel molecular descriptor known as a molecular quantum number (MQN).[62] This representation, inspired by the periodic table of the elements, assigns a molecule to a position in a multidimensional grid based on a set of 42 values calculated for each molecule. These values include counts of specific types of atoms and bonds, polarity, and topology. This grid enables rapid identification of closely related compounds and facilitates a high performance visualization.

Another recent method, which is somewhat similar to the SOM described above, is generative topographic mapping (GTM).[63] In GTM, a set of radial basis functions is used to reduce a set of high dimensional vectors representing molecules to a two-dimensional grid representation that can be visualized. Results published by Varnek and co-workers[63] have shown that a two-dimensional map produced by GTM more faithfully reproduces the set of distances in the original high-dimensional space. As a demonstration of the method, Lin and co-workers[64] generated a GTM from a subset of the GDB-17 database known as FDB-17. The FDB-17 database consists of 10 million lead-like and fragment-like molecules,

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124

which the authors claim represents the entirety of lead-like and fragment-like chemical space. The authors used fragment descriptions similar to the fingerprints described above and selected a set of reference "frame" molecules similar to the "satellite" molecules described by Oprea in the work discussed above. Once constructed, the GTMs provided 2D maps of chemical space that were used to compare databases and assess chemical space coverage. While it has been demonstrated that techniques like FUn and GTM can be used to visualize large virtual libraries, both methods will still encounter many of the same distortions associated with projections from high-dimensional spaces to a two or three-dimensional visualization.

While DNA encoded libraries are not technically virtual, their size, which can range into billions of molecules, can present challenges similar to those encountered with large virtual libraries. In a recent paper, Kontijevskis[65] describes a method for encoding chemical structures as sets of rings and linkers derived from a representation originally published by Bemis and Murcko.[66] In this representation, known as reduced complexity molecular frameworks (RCMF), a molecule is represented by a text string that provides unique identifiers for ring systems, linkers, and the angles between ring systems. Because of the generality of the representation, large numbers of molecules will map to the same RCMF. The author found that 99% of small "drug-like" compounds could be represented by 452 RCMF types. This reduced representation enables very large libraries to be mapped into a heatmap for comparison and diversity analysis. The maps generated from RMCF were subsequently used to compare the chemical space coverage of a number of DNA encoded libraries with the space covered by drug-like compounds in the ChEMBL and PubChem databases.

■ PRACTICAL CONSIDERATIONS

Virtual screening[67,68] has become a mainstay of modern drug discovery and is now commonly used in both academic and industrial medicinal chemistry. Over the past 25 years, researchers have developed a number of different virtual screening methods that vary in terms of the amount of information necessary and the time required to perform the calculations. Virtual screening methods can be roughly divided into two categories, ligand-based methods[69] and structure-based methods.[70] In ligand-based virtual screening one starts with a known molecule and searches a chemical library (either real or virtual) to identify similar molecules. Searches can be performed for molecules that are similar based on 2D topology or for molecules with similar 3D characteristics such as shape or arrangement of pharmacophoric features. Topological or 2D searches, which require only information on atom types and connectivity, are very fast and can often process thousands of molecules per second. Similarity searches involving 3D information are slower but can still be optimized to run on large databases. One drawback to ligand-based methods is that they require a known structure as input. In cases where we have a novel target with no known binders, ligand-based virtual screening has limited utility. In structure-based virtual screening, chemical libraries are searched for molecules that are compatible with a protein binding site. Since these methods require an exploration of ligand conformation as well as binding site compatibility, searches can take anywhere from a few seconds to a minute per molecule. Structure-based virtual screening has the advantage that it requires only a protein binding site and can be used when known binders do not exist. The primary drawback to structure-based virtual screening is its limited accuracy. Methods for scoring the interactions between small molecules and proteins are still somewhat crude, and hit rates from structure-based virtual screening can be low.

One of the prime motivators behind the construction of large virtual libraries has been the desire to improve the output of virtual screening methods. There is a common belief that the quality of hits from virtual screening can be improved by considering a larger source pool of molecules. While this may be true, there are confounding factors. When we consider a virtual screen, we can encounter two types of errors. A "false negative" is a screening hit, which is incorrectly assigned as inactive and is missed by the virtual screen. A "false positive" is an inactive molecule that is predicted to be active. As pointed out more than 20 years ago by Jain,[71] false positives are the largest problem in virtual screening. If we consider a virtual screen with a false positive rate of 1% (an optimistic estimate for even the best virtual screening methods), a virtual screen on a library of 1 million molecules would yield 10 000 false positive hits. The impact of 10 000 false positive hits would, of course, depend on the situation. In a screen of a large corporate collection the expense of formatting and screening the additional compounds might be acceptable. However, if the assay being run were expensive or if compound synthesis were required, the cost of the false positives could be prohibitive. We simply do not have enough experience with virtual screening on libraries with hundreds of millions of compounds to understand the true impact of false negative and false positive rates. It may be with very large virtual libraries we can identify the true hits in "long tail" of the score distribution. It may also be that the false positive hits will completely swamp out the signal from the true positive hits. Hopefully experience with these larger databases will enable us to further tune our virtual screening methods.

Another consideration with very large virtual libraries is the time and CPU resource required for processing. As an example, let us consider the time required to process the Enamine REAL database of 680 million compounds. As a first step, we will consider substructure and similarity searches. Since these searches only require a description of molecular connectivity and atom types, they are relatively fast. A brute force similarity search of a 680 million compound database would require several hours to complete, but clever indexing strategies can dramatically reduce the time requirements for these searches. A number of groups have recently developed indexing methods that enable typical searches of a 680 million compound database to be performed in a few seconds.[72,73]

Virtual screening methods such as protein−ligand docking and 3D similarity searches, which require an ensemble of three-dimensional coordinates as input, are considerably more time-consuming. The first step for most 3D virtual screening methods is the generation of an ensemble of 3D conformations. State of the art conformation generation programs can typically process 1−2 molecules per second on a single CPU core. Given a speed of 2 molecules/second, 3830 CPU days would be required to process a database of 680 million compounds. While this time estimate appears imposing, the process can be readily parallelized across multiple CPU cores. If the conformer generation step is split across 1000 CPU cores and run on cloud computing resources, it can be performed in less than 4 days. At the time of this writing, time on a 72 CPU computer can be purchased for

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124

$1.08/h. As such, the conformer generation would only cost $1380. Docking small molecules into a protein binding site is at least 2−20× more compute expensive than conformer generations. Even the fastest docking programs require 2 s per molecule to dock an ensemble of conformations into a protein binding site. At this rate, approximately 15 327 CPU days would be required to dock 680 million molecules. As with conformer generation, the docking process can be parallelized and run on cloud resources. Given the price estimates above, a database of 680 million compounds could be docked in 16 days on 1000 CPU cores at a cost of approximately $5551. However, although the cost of running these large virtual screens is relatively low, implementation of such workflows requires considerable expertise. Hopefully, as virtual screening methods evolve, the access to sufficient computational resources will become more straightforward.

CONCLUSIONS AND FUTURE DIRECTIONS

Over the past few years, the size of virtual chemical libraries has continued to grow. A few years ago, a typical virtual library consisted of a few million compounds. These libraries could be processed on computer workstations or local high performance computing (HPC) clusters. As can be seen above, we now have access to libraries that are rapidly approaching a billion compounds. We are reaching a point where the storage and processing of these libraries becomes a challenge, even for seasoned computational chemists. Beyond the requirements for computing resources and expertise, perhaps the most important question is whether the current generation of virtual screening methods are accurate enough to effectively process libraries containing hundreds of millions of molecules. As mentioned above, false positive hits are a major limitation for any virtual screening method. A virtual screening method with a false positive rate of even 1% will be of limited value when screening a database of 100 million compounds.

In order to more efficiently search the vastness of chemical space, we have to move beyond our current brute force approach of evaluating molecules one at a time. Newer models, which take advantage of advances in machine learning, may provide a means of navigating this space. If we can develop a continuous representation of chemical space, we can potentially exploit gradients in that space to identify optimal regions for exploration. This navigation will, of course, require higher quality predictive models than those currently in existence. Hopefully, continued advances in predictive methods will enable us to identify the needles in our increasingly growing haystack.

We have entered an exciting era where data science is becoming a key element in drug discovery. Large virtual libraries are just one component in the computational drug discovery workflow. While it is tempting to think that mountains of data alone will dramatically improve productivity, a concerted effort will be required to merge predictive models with the growing number of readily available molecules.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pwalters@relaytx.com.

ORCID

W. Patrick Walters: 0000-0003-2860-7958

Notes

The author declares no competing financial interest.

Biography

W. Patrick Walters heads the Computation & Informatics group at Relay Therapeutics in Cambridge, MA. His group focuses on novel applications of computational methods that integrate computer simulations and experimental data to provide insights that drive drug discovery programs. Prior to joining Relay, he spent more than 20 years at Vertex Pharmaceuticals where he was Global Head of Modeling & Informatics. Pat received his Ph.D. in Organic Chemistry from the University of Arizona where he studied the application of artificial intelligence in conformational analysis. Prior to obtaining his Ph.D., he worked at Varian Instruments as both a chemist and a software developer. Pat received his B.S. in Chemistry from the University of California, Santa Barbara.

ACKNOWLEDGMENTS

The author thanks Tushar Gupta, Nicholas Freitas, Brandi Hudson, Steven Kearnes, Demetri Moustakas, Mark Murcko, Levi Pierce, Molly Schmidt, and Jonathan Weiss for discussions of early versions of this manuscript.

ABBREVIATIONS USED

GDB, generated database; NMDA, *N*-methyl-D-aspartate; MQN, molecular quantum number; RCMF, reduced complexity molecular formula; nAChR, nicotinic acetylcholine receptor; CPU, central processing unit

REFERENCES

(1) Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-Driven Medicinal Chemistry in the Era of Big Data. *Drug Discovery Today* **2014**, *19*, 859−868.

(2) Richter, L.; Ecker, G. F. Medicinal Chemistry in the Era of Big Data. *Drug Discovery Today: Technol.* **2015**, *14*, 37−41.

(3) Tetko, I. V.; Engkvist, O.; Chen, H. Does "Big Data" Exist in Medicinal Chemistry, and if So, How Can It Be Harnessed? *Future Med. Chem.* **2016**, *8*, 1801−1806.

(4) https://www.prnewswire.com/news-releases/cas-assigns-the-100-millionth-cas-registry-number-to-a-substance-designed-to-treat-acute-myeloid-leukemia-300106332.html (accessed August 20, 2018).

(5) Weininger, D. Combinatorics of Small Molecular Structures. In *Encyclopedia of Computational Chemistry*; von Ragué Schleyer, P., Ed.; John Wiley & Sons, Ltd: Chichester, U.K., 2002; Vol. *8*, p 1056.

(6) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: a Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(7) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-Like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.

(8) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic Combinatorial Analysis Procedure: a Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

(9) Ogata, K.; Isomura, T.; Yamashita, H.; Kubodera, H. A Quantitative Approach to the Estimation of Chemical Space From a Given Geometry by the Combination of Atomic Species. *QSAR Comb. Sci.* **2007**, *26*, 596−607.

(10) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(11) Rishton, G. M. Reactive Compounds and in Vitro False Positives in HTS. *Drug Discovery Today* **1997**, *2*, 382−384.

(12) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-Like Chemical Space Based on GDB-17 Data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124

(13) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722−730.
(14) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe Up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.
(15) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(16) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(17) Nguyen, K. T.; Luethi, E.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J.-L. 3-(Aminomethyl)Piperazine-2,5-Dione as a Novel NMDA Glycine Site Inhibitor From the Chemical Universe Database GDB. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3832−3835.
(18) Nguyen, K. T.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J.-L. Discovery of NMDA Glycine Site Inhibitors From the Chemical Universe Database GDB. *ChemMedChem* **2008**, *3*, 1520−1524.
(19) Bréthous, L.; Garcia-Delgado, N.; Schwartz, J.; Bertrand, S.; Bertrand, D.; Reymond, J.-L. Synthesis and Nicotinic Receptor Activity of Chemical Space Analogues of N-(3 R)-1-Azabicyclo[2.2.2]Oct-3-Yl-4-Chlorobenzamide (PNU-282,987) and 1,4-Diazabicyclo[3.2.2]Nonane-4-Carboxylic Acid 4-Bromophenyl Ester (SSR180711). *J. Med. Chem.* **2012**, *55*, 4605−4618.
(20) Garcia-Delgado, N.; Bertrand, S.; Nguyen, K. T.; van Deursen, R.; Bertrand, D.; Reymond, J.-L. Exploring A7-Nicotinic Receptor Ligand Diversity by Scaffold Enumeration From the Chemical Universe Database GDB. *ACS Med. Chem. Lett.* **2010**, *1*, 422−426.
(21) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages Into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.
(22) van Deursen, R.; Reymond, J.-L. Chemical Space Travel. *ChemMedChem* **2007**, *2*, 636−640.
(23) Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A. Organic Synthesis Provides Opportunities to Transform Drug Discovery. *Nat. Chem.* **2018**, *10*, 383−394.
(24) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data From Pharmaceutical Patents: a Computational Analysis of Medicinal Chemists' Bread and Butter. *J. Med. Chem.* **2016**, *59*, 4385−4402.
(25) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59*, 4443−4458.
(26) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. LEAP into the Pfizer Global Virtual Library (PGVL) Space: Creation of Readily Synthesizable Design Ideas Automatically. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Methods in Molecular Biology, Vol. 685; Humana Press: Totowa, NJ, 2010; pp 253−276.
(27) Hu, Q.; Peng, Z.; Sutton, S. C.; Na, J.; Kostrowicki, J.; Yang, B.; Thacher, T.; Kong, X.; Mattaparti, S.; Zhou, J. Z.; Gonzalez, J.; Ramirez-Weinhouse, M.; Kuki, A. Pfizer Global Virtual Library (PGVL): a Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information. *ACS Comb. Sci.* **2012**, *14*, 579−589.
(28) Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016**, *56*, 1253−1266.
(29) Chevillard, F.; Kolb, P. SCUBIDOO: a Large Yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized Toward High Likelihood of Synthetic Tractability. *J. Chem. Inf. Model.* **2015**, *55*, 1824−1835.
(30) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for in Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51*, 3093−3098.
(31) Humbeck, L.; Weigang, S.; Schäfer, T.; Mutzel, P.; Koch, O. CH IPMUNK: a Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, Exploitable for Protein-Protein Interaction Modulators. *ChemMedChem* **2018**, *13*, 532−539.
(32) https://www.emolecules.com (accessed August 20, 2018).
(33) https://www.molport.com (accessed August 20, 2018).
(34) Schäfer, T.; Mutzel, P. StruClus: Scalable Structural Graph Set Clustering with Representative Sampling. In *Advanced Data Mining and Applications*; Lecture Notes in Computer Science, Vol. 10604; Springer International Publishing: Cham, Switzerland, 2017; pp 343−359, DOI: 10.1007/978-3-319-69179-4_24.
(35) https://enamine.net (accessed August 20, 2018).
(36) Schneider, G.; Fechner, U. Computer-Based De Novo Design of Drug-Like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.
(37) Schneider, G. Generative Models for Artificially-Intelligent Molecular Design. *Mol. Inf.* **2018**, *37*, 1880131.
(38) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(39) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 191.
(40) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.
(41) Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A. Adversarial Threshold Neural Computer for Molecular De Novo Design. *Mol. Pharmaceutics* [Online early access]. DOI: 10.1021/acs.molpharmaceut.7b01137. Published Online: March 23, **2018**. https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.7b01137 (accessed August 20, 2018).
(42) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De NovoDesign of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37* (1−2), 1700153−1700154.
(43) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(44) Bonnet, P. Is Chemical Synthetic Accessibility Computationally Predictable for Drug and Lead-Like Molecules? a Comparative Assessment Between Medicinal and Computational Chemists. *Eur. J. Med. Chem.* **2012**, *54*, 679−689.
(45) Baber, J.; Feher, M. Predicting Synthetic Accessibility: Application in Drug Discovery and Development. *Mini-Rev. Med. Chem.* **2004**, *4*, 681−692.
(46) Huang, Q.; Li, L.-L.; Yang, S.-Y. RASA: a Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-Like Molecules. *J. Chem. Inf. Model.* **2011**, *51*, 2768−2777.
(47) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311−325.
(48) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.
(49) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.
(50) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103−1113.
(51) Xue, L.; Bajorath, J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124

Screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363−372.
(52) Yap, C. W. PaDEL-Descriptor: an Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.
(53) Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold 2, Molecular Descriptors From 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337−1344.
(54) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(55) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58−63.
(56) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(57) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a Preferred Set of Molecular Descriptors for Compound Classification Based on Principal Component Analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699−704.
(58) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? a Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49*, 108−119.
(59) Schneider, P.; Tanrikulu, Y.; Schneider, G. Self-Organizing Maps in Drug Discovery: Compound Library Design, Scaffold-Hopping, Repurposing. *Curr. Med. Chem.* **2009**, *16*, 258−266.
(60) Yan, A. Application of Self-Organizing Maps in Compounds Pattern Recognition and Combinatorial Library Design. *Comb. Chem. High Throughput Screening* **2006**, *9*, 473−480.
(61) Probst, D.; Reymond, J.-L. FUn: a Framework for Interactive Visualizations of Large, High Dimensional Datasets on the Web. *Bioinformatics* **2018**, *34*, 1433−1435.
(62) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **2009**, *4*, 1803−1805.
(63) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* **2012**, *31*, 301−312.
(64) Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J.-L.; Varnek, A. Mapping of the Available Chemical Space Versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018**, *13*, 540−554.
(65) Kontijevskis, A. Mapping of Drug-Like Chemical Universe with Reduced Complexity Molecular Frameworks. *J. Chem. Inf. Model.* **2017**, *57*, 680−699.
(66) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(67) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862−865.
(68) Klebe, G. Virtual Ligand Screening: Strategies, Perspectives and Limitations. *Drug Discovery Today* **2006**, *11*, 580−594.
(69) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-Art in Ligand-Based Virtual Screening. *Drug Discovery Today* **2011**, *16*, 372−376.
(70) Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *AAPS J.* **2012**, *14*, 133−141.
(71) Jain, A. N. Scoring Noncovalent Protein-Ligand Interactions: a Continuous Differentiable Function Tuned to Compute Binding Affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.
(72) https://chemaxon.com/products/madfast (accessed August 20, 2018).
(73) https://www.slideshare.net/NextMoveSoftware/recent-advances-in-chemical-biological-search-systems-evolution-vs-revolution (accessed August 20, 2018).
(74) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: an Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.
(75) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633.
(76) Li, Q.; Cheng, T.; Wang, Y.; Bryant, S. H. PubChem as a Public Resource for Drug Discovery. *Drug Discovery Today* **2010**, *15*, 1052−1057.
(77) Pence, H. E.; Williams, A. ChemSpider: an Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87*, 1123−1124.
(78) Williams, A. J.; Tkachenko, V.; Golotvin, S.; Kidd, R.; McCann, G. ChemSpider - Building a Foundation for the Semantic Web by Hosting a Crowd Sourced Databasing Platform for Chemistry. *J. Cheminf.* **2010**, *2*, O16−1.
(79) Irwin, J. J.; Shoichet, B. K. ZINC−a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(80) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: a Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2016**, *44*, D1220−D1228.

DOI: 10.1021/acs.jmedchem.8b01048
J. Med. Chem. 2019, 62, 1116−1124