JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Cite This: J. Chem. Inf. Model. 2019, 59, 644−651

Perspective

pubs.acs.org/jcim

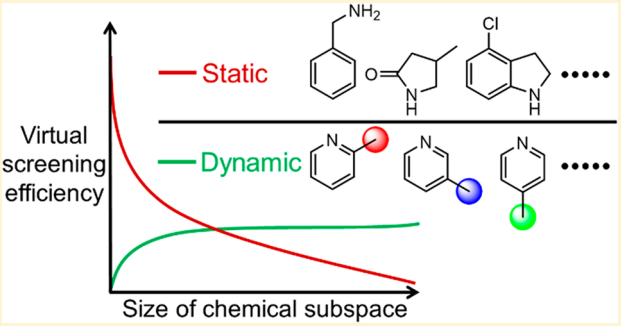
Virtual Compound Libraries in Computer-Assisted Drug Discovery

Niek van Hilten,[†] Florent Chevillard, and Peter Kolb*

Department of Pharmaceutical Chemistry, Philipps-University Marburg, Marbacher Weg 6, 35032 Marburg, Germany

Supporting Information

ABSTRACT: The use of virtual compound libraries in computer-assisted drug discovery has gained in popularity and has already lead to numerous successes. Here, we examine key static and dynamic virtual library concepts that have been developed over the past decade. To facilitate the search for new drugs in the vastness of chemical space, there are still several hurdles to overcome, including the current difficulties in screening and parsing efficiency and the need for more reliable vendors and accurate synthesis prediction tools. These challenges should be tackled by both the developers of virtual libraries and by their users, in order for the exploration of chemical space to live up to its potential.



In drug design, the traditional way to find lead candidates is by high-throughput *in vitro* testing of large ($\sim 10^6$–10^7) physically existing compound libraries for interaction with the target of interest. After the "rise and fall of combinatorial chemistry",[1] medicinal chemists increasingly focused on virtual compound libraries that are stored on computer hard-drives rather than pharmaceutical company shelves. In order to improve time- and cost-efficiency, virtual screening can be performed with those virtual compound libraries to narrow down the number of molecules to evaluate as potential lead candidates, prior to synthesis and experimental testing.

The total number of possible unique organic molecules, referred to as chemical space, is immense. Estimates range from 10^{23} to an even more incomprehensible 10^{180}, depending on the inclusion of criteria such as the number of heavy atoms, atom types, ring strain, and stereochemistry.[2,3] Without doubt, this wealth of virtual chemicals holds many bioactive compounds and undiscovered therapeutic agents. The question is how to handle and navigate the ever increasing chemical space in order to find these precious few molecules.

In the past decade, designing drug-like compounds from smaller fragments has gained in popularity.[4,5] In fragment-based drug discovery (FBDD), small molecule fragments (MW ≤ 300 Da) are screened virtually or experimentally[6] and, if successful, expanded through growing, linking or merging strategies[7,8] to further improve binding to their protein target(s). Due to their low molecular weight and often favorable ADMET properties, fragments make excellent starting points for *de novo* drug design. Moreover, because of their compactness and few polar interaction sites, fragments are more likely than drug-like compounds to interact with a greater number of proteins, potentially yielding higher hit rates.[9] FDA approval of Vemurafenib[10] (a B-Raf kinase inhibitor) and Venetoclax[11] (an inhibitor for the apoptosis regulator BCL-2) mark important successes for this field and underline the potential of FBDD.

Although most fragment-based subsets of chemical space are smaller and more manageable than drug-like collections, several challenges remain. How does one select subsets that are big and diverse enough to be likely to contain potential hits, while preserving screenable proportions? How does one guarantee the possibility to chemically expand the fragment without breaking favorable interactions with the protein target? And, how does one ensure synthetic accessibility of the compounds resulting from such a virtual screening? Here, we examine the state-of-the-art of both static and dynamic virtual compound libraries that were recently developed to answer these questions (see Table 1). It is important to note that this paper is not aimed at being exhaustive. Rather, our goal is to provide the reader with concepts and some striking examples to help them make informed decisions when developing or using virtual libraries in their research.

STATIC VIRTUAL LIBRARIES

First, and perhaps most intuitive, there are static libraries of all unique virtual molecules that may exist within a certain set of boundaries. Currently, the most advanced example is GDB-17, which contains 116 billion virtual compounds with up to 17 atoms of C, N, O, S, and halogens within basic chemical rules.[12,13]

Although being the largest enumerated virtual compound library to date, GDB-17 still accounts for only a tiny fraction of drug-like chemical space. Extrapolation to 38 heavy atoms by Polishchuk et al.[14] shows that chemical space within the GDB boundaries and MW ≤ 500 Da for drug-likeness[15] contains

Received: October 22, 2018
Published: January 9, 2019

DOI: 10.1021/acs.jcim.8b00737
J. Chem. Inf. Model. 2019, 59, 644−651

Table 1. Overview of Some Key Static and Dynamic Virtual Libraries[a]

name	size	compound availability	construction method	comments and utilities	ref
			Static Virtual Libraries		
GDB-17	~166×10^9	–	all possible molecules of up to 17 atoms of C, N, O, S and halogens	freely accessible specifically aimed at FBDD	12
ZINC	~22×10^6	purchasable	collection of compounds commercially available at several vendors	freely accessible, drug-like subsets available, purchasability often not up-to-date	20
Enamine REAL	>300×10^6	purchasable	collection of compounds commercially available at Enamine	freely accessible	
SCUBIDOO	~21×10^6	synthesizable	7805 commercially available building blocks recombined with 58 reactions	freely accessible small, medium, large subsets available	24
SAVI	~283×10^6	synthesizable	products from robust chemistry on building blocks available from Sigma-Aldrich	freely accessible	25
CHIPMUNK	>95×10^6	synthesizable	products from robust chemistry on building blocks from ZINC, eMolecules, and Molport	freely accessible	26
			Dynamic Virtual Libraries		
Ftrees-FS	>10^{18}	synthesizable	16k retrosynthesis-based building blocks recombined with 11 reactions; up to 5 fragments per molecule	on-the-fly similarity searching	27
PINGUI	~$10-10^4$ (per query)	synthesizable	user-defined retrosynthesis and forward synthesis tools based on 58 robust organic reactions	freely accessible, fragment growing and SAR studies	7
FINDERS and REACT2D	~40×10^6	synthesizable	FINDERS identifies building blocks that are compatible with a user-defined reaction	freely accessible, 56 reactions are implemented, users can sketch additional reaction schemes, demonstrated with a library of ~100 000 purchasable molecules	28
AUTOCOUPLE	~10^4-10^5	synthesizable	couples commercially available building blocks to a user-defined fragment in a one-step virtual synthesis	demonstrated with a library of ~240 000 purchasable building blocks to grow a CREB-binding protein ligand fragment	29
Nikitin et al.	>10^{13}	synthesizable	constructed from ~400 combinatorial libraries, extended with suitable R-groups	searched by an in-house *de novo* drug design program	30
AllChem	>10^{20}	synthesizable	~7000 building blocks recombined with ~100 reactions; up to 3 fragments per molecule	3D similarity searches; provides synthetic routes	31
BI-CLAIM	~5×10^{11}	synthesizable	based on in-house combinatorial libraries and ~30k building blocks	similarity searches using Ftrees-FS	32
PGVL	~$10^{14}-10^{18}$	synthesizable	1244 well-characterized reaction applied on Pfizer in-house and ACD building blocks	similarity searches using Ftrees-FS,[36] atom-pair fingerprints,[37] and LEAP1/2[33]	33
PLC	~10^{11}	synthesizable	10 well-characterized reactions applied on in-house building blocks	synthesis can be automated using the Eli Lilly synthesis robot	35

[a]More details are in the main text. See Table S1 for all URLs (if available).

DOI: 10.1021/acs.jcim.8b00737
J. Chem. Inf. Model. 2019, 59, 644−651

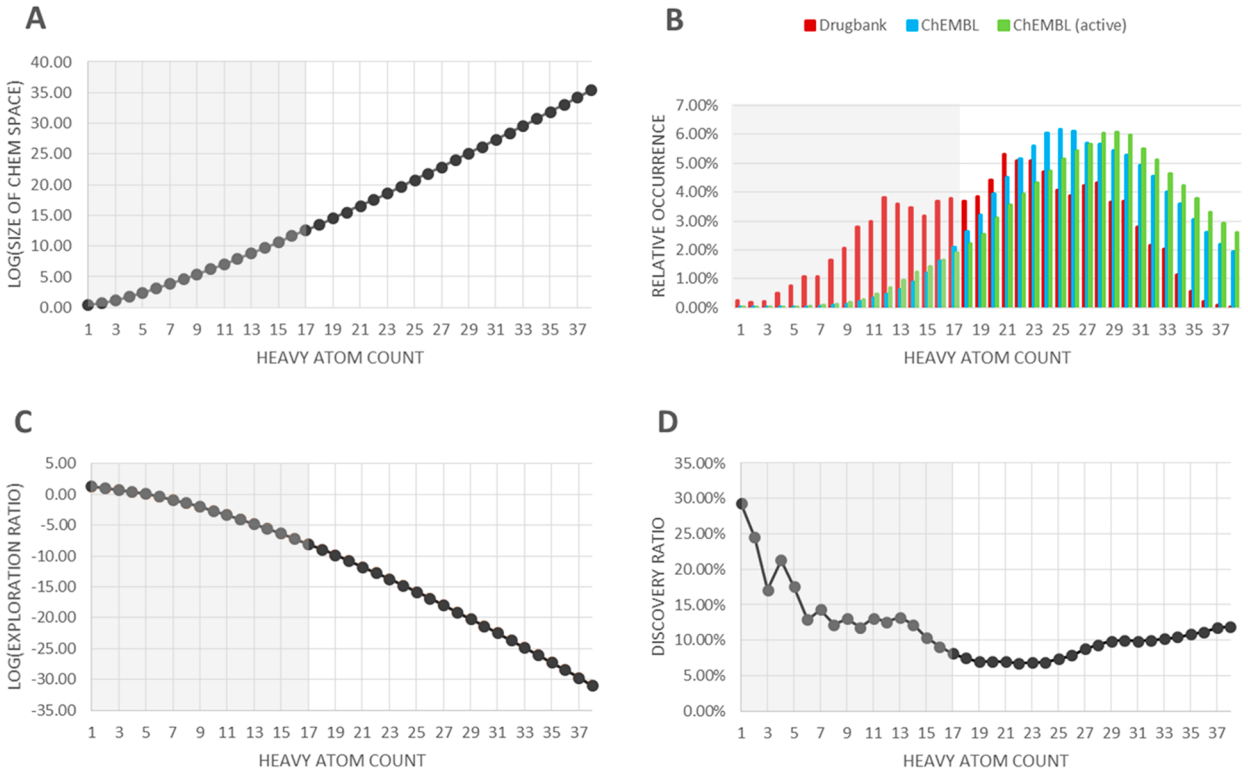


Figure 1. (A) Relation between the number of heavy atoms and the size of chemical space, according to the construction rules for GDB-17, as estimated by Polishchuk et al.[14] (B) Relative frequencies of heavy atom counts in the drug-like (MW ≤ 500 Da) portion of the DrugBank database (red bars; 6176 compounds), the ChEMBL database (blue bars; 1592286 compounds), and the active molecules in the ChEMBL database (green bars 140509 compounds). (C) Exploration ratio for each heavy atom count. (D) Discovery ratio for each heavy atom count. In all figures, the coverage of GDB-17 is indicated by a gray box. Data was processed using RDKit.[19] See the SI for details on the data analysis.

approximately 10^{33} molecules; a number that is 10^{22} times larger than the size of GDB-17 (Figure 1A).

Despite this quite limited coverage due to the sheer enormity of chemical space, fragment-like molecules of up to 17 heavy atoms (MW ≤ 300 Da) already make up approximately 35% of the approved drug-like molecules in the DrugBank database[16] (Figure 1B). Interestingly, this bias toward very small fragment-like molecules is less apparent in the activity data in the ChEMBL database.[17] In fact, the weighted average number of heavy atoms is slightly bigger for active compounds in ChEMBL (with a K_i-value for at least one target; data processed as described by Kramer et al.[18]) compared to the entire ChEMBL database: 26.92 versus 26.31 heavy atoms (Figure 1B).

When we look at the exploration ratio (the number of compounds in ChEMBL divided by the number estimated in Polishchuk's GDB-17 extrapolation of chemical space), it becomes clear that even within the GDB-17 range (i.e., ≤17 heavy atoms), many compounds are still unexplored. For example, for 13 heavy atoms, only one molecule is present in ChEMBL for (roughly) every 10^5 compounds in GDB-17 and for 17 heavy atoms this number is already at 10^8 (Figure 1C). This means that even within the rules of GDB-17, a large portion of fragment-like chemical space remains unexplored. At the same time, the discovery ratio (the number of active compounds in ChEMBL divided by the total in ChEMBL) in this regime is actually quite high: almost 15% of the 13 heavy atom molecules in ChEMBL have a reported activity for at least one target and for 17 heavy atom compounds this discovery ratio is still more than 8% (Figure 1D).

Taken together, these numbers illustrate that with respect to the tiny fraction of theoretical chemical space they span, smaller molecules (fragments) are overrepresented among approved drugs. Moreover, even within the <17 heavy atom range, millions of molecules remain unexplored. Assuming that the trend in Figure 1D holds true for these unexplored entities as well, a very significant ∼10% of those can be expected to have biological activity. On top of this aspect, fragments are not only valuable by themselves but often serve as starting points for further derivatization and expansion (via growing, linking or merging) as well, demonstrating the enormous potential behind a fragment-based approach.

On the downside, GDB-17's sheer size restricts the user to fast two-dimensional ligand-based techniques, such as computationally inexpensive similarity searches. In this respect, further expansion of the GDB, while in principle desirable to increase coverage, would only make matters worse. Also, should such a ligand-based screen yield a promising hit candidate, its synthetic preparation might be challenging, despite strict rules implemented by the creators. Therefore, the advantage of enumerating every possible molecule needs to be balanced with the potential problems in terms of searching and synthesizing molecules. So, what other aspects can be taken into account to focus libraries for different applications?

In every drug development project, the possibility to purchase or synthesize a virtual lead candidate for experimental testing is crucial. The ZINC database originally focused on commercially available, ready-made compounds (over 22 million, currently).[20] Due to its relatively modest size, the provision of 3D conformations, and free accessibility, ZINC is

DOI: 10.1021/acs.jcim.8b00737
J. Chem. Inf. Model. 2019, 59, 644−651

commonly used in many ligand- and structure-based virtual screening studies.[21,22] In practice, however, the purchasability of a compound can be unreliable, due to inconsistent and changeable vendor inventories.

Some vendors have also published downloadable databases based on their own in-stock molecules, as exemplified by Enamine (see Table S1 for URL). Arguably, such a database's content is more readily kept up-to-date, as it only contains molecules from one vendor inventory. Enamine has increasingly included make-on-demand compounds, for which they estimate a high likelihood of synthesizability (named the REAL database). These molecules are also largely part of the ZINC database and can thus be screened just like other molecules.

Alternatively, synthetic feasibility may be predicted by only enumerating molecules that originate from robust chemical reactions applied to commonly available building blocks, as, for example, proposed by Hartenfeller et al.[23] We used this concept to develop SCUBIDOO, a freely accessible database that contains 21 million virtual compounds, a number that is still feasible for structure-based computations.[24] Moreover, we used stratified balanced sampling to generate smaller representative subsets in order to screen the database more efficiently. What should be taken into account here is that the makeup of the initial set of building blocks heavily influences the resulting virtual compound library, because the computational synthesis steps rely on the compatibility with the building blocks. For example, when generating the SCUBIDOO database, application of only four of Hartenfeller's 58 chemical reactions already accounted for 75% of the resulting virtual compounds and 13 reactions were not applied at all due to building block incompatibility.

On a somewhat larger scale, but according to the same principle, the Synthetically Accessible Virtual Inventory (SAVI) database puts together highly annotated building blocks from Sigma-Aldrich to generate over 283 million products.[25] Like SCUBIDOO, SAVI is freely available online and provides the user with a synthetic robustness score and specific chemical warnings that could apply.

Within the same line of thought, Humbeck et al. recently developed a virtual library called CHIPMUNK.[26] The 95 million virtual compounds in this library are derived from educts of ZINC, eMolecules, and MolPort that are recombined *in silico* in three robust reaction categories, aiming to ensure synthesizability. Besides covering drug-like protein−ligand inhibitor space, the authors aimed to go beyond Lipinski's rule of five to include the physicochemical profiles of protein−protein interaction inhibitors (PPIIs) as well.

In general, enumerating all (synthesizable) chemicals and filtering or screening them *post hoc* is an intuitive way of handling large virtual compound spaces. It also allows for analysis of the contents of chemical space and whether we are making the right molecules at all. One could look at static virtual compounds libraries as giant haystacks which can be further pruned to smaller haystacks based on a user's needs (e.g., filtering according to the Lipinski rules, ADMET, solubility, price, availability, etc.). Preparation of the libraries for virtual screening (3D, protonation, or stereoisomer generation) should, in our opinion, also be left in the hands of the users, as data preparation protocols depend on which software one plans to use.

Finally, it is important to realize that a static approach will always confront the user with a trade-off between completeness and screenability: (1) a complete virtual library is not easily screenable with current basic hardware and software configurations and (2) a screenable virtual library is far from complete (in the case of e.g. SCUBIDOO due to a bias introduced by the set of reactions and building blocks).

DYNAMIC VIRTUAL LIBRARIES

One can bypass the aforementioned issues by only generating the chemical space around the virtual compounds that are relevant to the user's research question in the first place. Such dynamic virtual libraries allow for more efficient virtual screening, since full enumeration is not required. Often, the emphasis is on assuring synthetic feasibility. This can be done in two ways: (1) by using building blocks that result from the retrosynthesis of existing molecules via robust chemical reactions or (2) by simulating well-established chemistry used in combinatorial libraries (often within pharmaceutical companies).

An early example of the former approach was presented with the Ftrees-Fragment Spaces (Ftrees-FS) algorithm by Rarey et al.[27] This algorithm has been designed to perform similarity searches in a virtual compound space based on a set of building blocks that resulted from retrosynthesis of ∼30000 drug-like molecules using the RECAP approach. Up to five of those fragments are then recombined in a sequence of 11 one-step robust chemical reactions theoretically covering a 10^{18}-sized virtual compound space, without full enumeration.

Similarly, we have developed PINGUI,[7] a freely accessible online tool that uses the aforementioned 58 chemical reactions by Hartenfeller et al. to "deconstruct" a query molecule by identifying the most likely reactive site. In a next step, resulting fragments can be recombined with a set of user-defined compatible building blocks using the same types of chemistry, aiming to grow the core fragment to complement the chemical features of the target's binding site. Because the PINGUI tool couples only two rather than five fragments, the resulting chemical space is much more manageable than the one searched by the Ftrees-FS approach. Applying PINGUI to the design of β_2-adrenergic receptor ligands already showed its potential for reliable synthesis and improving binding affinity.[7]

Also based on the work by Hartenfeller et al., Pottel and Moitessier have developed FINDERS to identify chemicals that are compatible with a tool called REACT2D, which carries out the corresponding computational chemistry to create query-specific virtual libraries.[28] Free licenses for FINDERS and REACT2D are available to academia.

Within the same philosophy, Batiste et al. released AutoCouple, which takes commercially available building blocks and couples them in a one-step virtual organic reaction to a query molecule.[29] The "growing" of this user-defined query fragment is a bit more strict than the work based on Hartenfeller's reactions, as all virtual reaction products that contain an undesired reactive group (that thus would require additional protection to prevent cross reactivity) are excluded. By combining this software with molecular docking, the authors managed to find novel nanomolar ligands for their protein of interest (the CREB-binding protein bromodomain).

A more traditional way to ensure synthetic feasibility is by using the well-characterized chemistry of combinatorial libraries, which especially pharmaceutical companies often have extensive experience with. Initially, this revolutionary technique to synthesize large numbers of compounds at once was met with very high expectations in the medicinal chemistry field. However, when researchers realized that covering the

DOI: 10.1021/acs.jcim.8b00737
J. Chem. Inf. Model. 2019, 59, 644−651

vastness of chemical space and then retrieving the few bioactive compounds of interest was hardly feasible, combinatorial chemistry rapidly lost in popularity and was largely replaced by more targeted approaches.[1] In parallel to this trend, computational methods were developed for virtual combinatorial synthesis as well.

One of the first efforts in this category comes from Nikitin et al., who constructed a virtual chemical space of about 10^{13} in size by combining ∼400 published combinatorial libraries with commercially available reactants.[30] From 2005 onward, groups from different pharmaceutical companies developed such applications, tailor-made to their own in-house building block libraries and reactions: AllChem[31] by Tripos, CLAIM[32] by Boehringer Ingelheim (BI), and the Pfizer Global Virtual Library (PGVL[33]). In terms of synthetic accessibility, the developers of the Proximal Lilly Collection (PLC) are taking it a step further: their aim is to ensure that every compound in their 10^{11}-sized virtual library can be produced in an automated fashion by the robotic synthesis tools at Eli Lilly[34] using available in-house building blocks.[35]

Unfortunately, apart from the initial publications and some additional similarity search studies in the PGVL,[33,36,37] information on the impact of these applications on drug development (let alone usage of the tools themselves) is mostly not publicly available. It can be expected that during the development of most drugs, at least a focused virtual library has been created, particularly for those cases where the initial hit was fragment-sized. However, to the best of our knowledge, no case has been described where the original hit molecule stemmed from a virtual library.

CHALLENGES

The wealth of static and dynamic techniques to navigate chemical space offers great opportunities for medicinal chemists in search for novel bioactive molecules. However, several challenges remain for the developers of virtual libraries, as well as for their users.

First, perhaps the most obvious hurdle is screening efficiency. As mentioned before, using virtual libraries often confronts the user with the dilemma of completeness versus screenability. Put simply; the bigger the chemical subspace, the harder it is to handle and navigate. For developers, this challenge boils down to (1) developing new algorithms for more efficient parsing, (2) smart data organization or (3) reducing library size, while retaining molecules of interest (i.e., the true positives). In other words, for the efficiency to increase, virtual libraries should become more targeted toward the scientific questions asked by the user. This trend shows great similarity with what happened with combinatorial chemistry over the past two decades. In a static virtual library setting, stratified sampling has already been shown to be a promising method for the generation of small representative subsets.[24] If screening such a subset yields a virtual hit, the cluster that was represented by that compound can be taken into closer consideration in a follow-up effort. Intrinsically, dynamic virtual libraries already have a higher efficiency, since they only span a chemical subspace around a user-defined query molecule or fragment. A potential pitfall here could be that one is less likely to find completely novel scaffolds, as the library by definition is defined by the characteristics of the initial molecule. For the purpose of fragment growing, however, this approach has already been shown to be very useful: if part A of a molecule is known to be successful in binding to a certain region of the protein target, one could try to recombine it with different parts B to also target an adjacent region in the protein binding site.[7,8,38]

Second, and as a corollary to the previous point: how big can virtual libraries become before we run into storage problems? The 166 billion molecules of the GDB-17 stored as SMILES occupy "more than 400 GB in gzip format".[13] This leads to an estimated storage requirement of 2×10^{39} ZB (zettabyte) for all the 10^{60} possible drug-like molecules. Humankind currently generates 2.5 EB (exabyte) per day[39] or about 1 ZB per year, which already leads to an increasing gap in what we produce and what can be stored. Thus, if everyone stopped producing the stereotypical cat videos (and everything else) for an entire year and donated all the storage space to virtual libraries, we could store approximately 4×10^{20} virtual molecules. Until the invention of novel storage technologies that allow higher densities, this will be the upper limit of what chemoinformaticians can achieve (and, in the real world, far less). An aggravating factor is stereoisomer generation: even if virtual compound libraries are generated without explicit stereoinformation, the user has to generate stereoisomers before any virtual screening campaign. This will increase storage requirements by up to 2-fold. For instance, in the case of SCUBIDOO, 16 million virtual products were created at first, which inflated to a total of 21 million virtual products (roughly a 30% size increase) upon stereoisomer generation.[24]

Third, higher accuracy in the prediction of synthetic feasibility should assist the decision making of which molecule to pursue. Courtesy of the recent advances in deep learning methods and artificial intelligence combined with the knowledge of reactions extracted from the literature, it is now possible to rapidly evaluate whether a compound is synthesizable but also suggest an accurate synthetic route.[40−45] Also, such machine learning techniques can be used to translate chemical representations (like SMILES) into low-level encodings.[46,47] Next, one can efficiently generate a chemical subspace around a query molecule by introducing small perturbations to the compressed representations in the latent space and subsequently decoding these vectors back to SMILES format. We foresee that such models will be increasingly implemented within the next iterations of the current and future virtual compounds libraries.

Fourth, availability is an aspect the field as a whole could very much improve on. Unsurprisingly, many of the aforementioned virtual libraries are not, and will never be, available to people outside of the respective pharmaceutical companies in which they were developed. For the field of computational drug design to progress, however, it is important that such tools are freely accessible, at least for academia. In those cases where the libraries are available without restriction, a potential blind spot for virtual library developers is the importance of intuitive (online) interfaces that help the user to apply the tools to their problems as efficiently as possible. In the long run, availability alone is also not enough: libraries need to be maintained and supported to ensure their accessibility over several years. There are examples where databases have existed for more than a decade (e.g., ZINC), but, especially in the academic context, many databases (molecule or otherwise) die with the graduation of the student who developed them. It is beyond the scope of this perspective to suggest solutions to this problem, as it requires a community effort. At least in our lab, we ask ourselves carefully in each case, whether a new database is necessary, or whether

DOI: 10.1021/acs.jcim.8b00737
J. Chem. Inf. Model. 2019, 59, 644−651

we can include the data in some existing—and well-maintained—repository.

Lastly, assuring commercial or synthetic availability of the compounds in a virtual library remains challenging. In most current virtual libraries, developers attempt to tackle this issue by using computer encoded chemical synthesis routes that are well-established in the lab. Generally, this approach seems to be more successful than assuring commercial availability, since organic chemistry often turns out to be more reliable than vendor inventories. For this to stay true, however, continuous updating and close collaboration with organic chemists (e.g., the users) is required. One way to achieve this could be by explicitly asking virtual library users to submit feedback on whether their virtual screening-inspired synthesis campaigns were successful or not and adjust the library accordingly. Although some programs already started to implement this, a more realistic representation of the chemical reactions, e.g. including protecting group steps, a focus on building blocks free of interfering functional groups, and correct assignment and generation of stereochemistry, are necessary steps toward increased synthetic likelihood.

CONCLUSIONS

Chemical space is vast and computational techniques have become indispensable to navigate it. Arguably, static virtual libraries are the most intuitive way of handling big numbers of molecules. However, their sheer sizes drastically complicate screening efficiency. Therefore, there is a need for extracting relevant subsets. Since transition from virtual molecules to physical substances is crucial during a drug development campaign, a common approach is to enumerate synthesizable or commercially available compounds only.

Alternatively, dynamic virtual libraries provide the user with a concept or algorithm rather than an enumerated list of molecules. Almost exclusively, such techniques are based on on-the-fly virtual synthesis of chemical structures from a set of (fragment-like) commercially available building blocks. Parallel to the trend we have seen for combinatorial chemistry, such dynamic databases are increasingly becoming more targeted toward a user-defined problem.

Both static and dynamic virtual libraries should be used in synergy in a drug discovery effort. One could start by virtually screening representative subsets of large static libraries. Once promising hits have been identified with their associated synthetic route (building blocks and organic reaction), one could use this information to dynamically generate more compounds around the initial hits. Dynamic libraries allow us to easily iterate to improve the initial promising hits. A possible workflow to achieve this is shown in Figure 2.

Although virtual libraries already serve as valuable tools for medicinal chemists, several challenges remain for their developers and users. The field would greatly benefit from increased screening efficiencies, better accessibility to scientists outside of pharmaceutical industry and more reliable synthetic or commercial availability.

But maybe such libraries do not need to contain everything. One could argue that even the biggest libraries are still sparse (Figure 1C). Hence, we can only expect initial hit compounds from them, but not potent leads or even drugs. The development of such molecules will still require small changes to molecule structures with "traditional" medicinal chemistry. Thus, virtual libraries might serve as sources for initial hits and the establishment of structure−activity relationships (SAR).

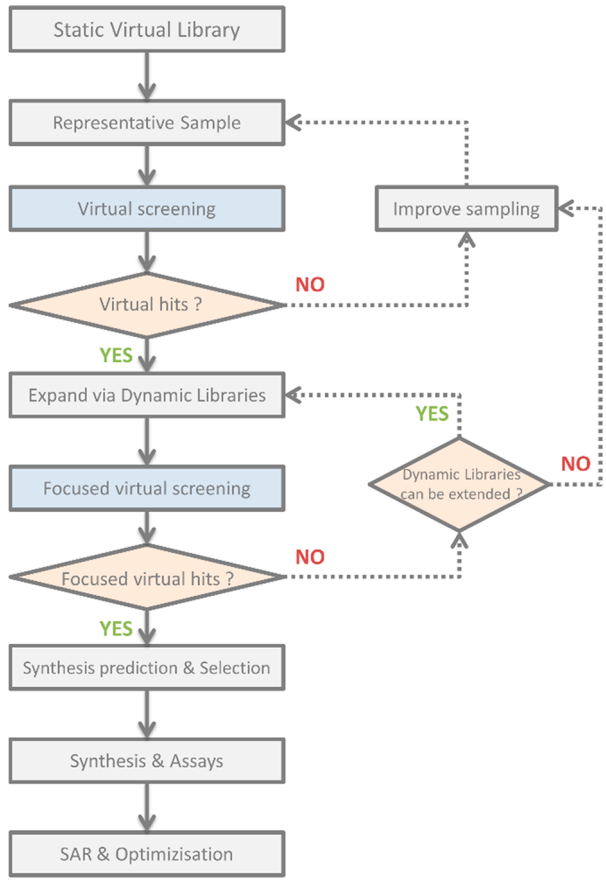


Figure 2. Possible workflow of the application of static and dynamic virtual compound libraries in a computer-assisted drug discovery project.

The most potent compounds, potentially analyzed with machine learning retrosynthesis tools for alternative synthesis pathways, can then be optimized further. In this way, as a joint effort, we should be able to make *in silico* exploration of chemical space live up to its potential to revolutionize the development of future medication.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00737.

Methodological details on the DrugBank and ChEMBL data analysis and the URLs of all mentioned virtual compound libraries (PDF)

AUTHOR INFORMATION

Corresponding Author

*Email: peter.kolb@uni-marburg.de.

ORCID

Peter Kolb: 0000-0003-4089-614X

Present Address

†Supramolecular and Biomaterials Chemistry, Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands.

Funding

P.K. thanks the German Research Foundation DFG for Heisenberg Professorship KO4095/4-1 and Emmy Noether

DOI: 10.1021/acs.jcim.8b00737
J. Chem. Inf. Model. 2019, 59, 644−651

fellowship KO4095/1-1. N.v.H. was supported through the Erasmus program of Radboud University, Nijmegen, and a scholarship granted by Duitsland Instituut Amsterdam.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Thomas J. Boltje at Radboud University, Nijmegen, for his feedback and support. We also thank the anonymous reviewers for their many excellent suggestions for additional analysis.

■ ABBREVIATIONS

FBDD, fragment-based drug discovery; MW, molecular weight; ADMET, absoprtion, distribution, metabolism, excretion, toxicity; FDA, food and drug administration; BCL-2, B-cell lymphoma 2; PPII, protein−protein interaction inhibitors; CREB, cAMP response element-binding; SAR, structure−activity relationships.

■ REFERENCES

(1) Kodadek, T. The Rise, Fall and Reinvention of Combinatorial Chemistry. *Chem. Commun. (Cambridge, U. K.)* **2011**, *47*, 9757−9763.
(2) 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.
(3) Gorse, A. D. Diversity in Medicinal Chemistry Space. *Curr. Top. Med. Chem.* **2006**, *6*, 3−18.
(4) Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. Twenty Years On: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug Discovery* **2016**, *15*, 605−619.
(5) Murray, C. W.; Verdonk, M. L.; Rees, D. C. Experiences in Fragment-Based Drug Discovery. *Trends Pharmacol. Sci.* **2012**, *33*, 224−232.
(6) Schiebel, J.; Krimmer, S. G.; Rower, K.; Knorlein, A.; Wang, X.; Park, A. Y.; Stieler, M.; Ehrmann, F. R.; Fu, K.; Radeva, N.; Krug, M.; Huschmann, F. U.; Glockner, S.; Weiss, M. S.; Mueller, U.; Klebe, G.; Heine, A. High-Throughput Crystallography: Reliable and Efficient Identification of Fragment Hits. *Structure* **2016**, *24*, 1398−1409.
(7) Chevillard, F.; Rimmer, H.; Betti, C.; Pardon, E.; Ballet, S.; van Hilten, N.; Steyaert, J.; Diederich, W. E.; Kolb, P. Binding-Site Compatible Fragment Growing Applied to the Design of Beta2-Adrenergic Receptor Ligands. *J. Med. Chem.* **2018**, *61*, 1118−1129.
(8) Hung, A. W.; Silvestre, H. L.; Wen, S.; Ciulli, A.; Blundell, T. L.; Abell, C. Application of Fragment Growing and Fragment Linking to the Discovery of Inhibitors of Mycobacterium Tuberculosis Pantothenate Synthetase. *Angew. Chem., Int. Ed.* **2009**, *48*, 8452−8456.
(9) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.
(10) Tsai, J.; Lee, J. T.; Wang, W.; Zhang, J.; Cho, H.; Mamo, S.; Bremer, R.; Gillette, S.; Kong, J.; Haass, N. K.; Sproesser, K.; Li, L.; Smalley, K. S.; Fong, D.; Zhu, Y. L.; Marimuthu, A.; Nguyen, H.; Lam, B.; Liu, J.; Cheung, I.; Rice, J.; Suzuki, Y.; Luu, C.; Settachatgul, C.; Shellooe, R.; Cantwell, J.; Kim, S. H.; Schlessinger, J.; Zhang, K. Y.; West, B. L.; Powell, B.; Habets, G.; Zhang, C.; Ibrahim, P. N.; Hirth, P.; Artis, D. R.; Herlyn, M.; Bollag, G. Discovery of a Selective Inhibitor of Oncogenic B-Raf Kinase with Potent Antimelanoma Activity. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 3041−3046.
(11) Souers, A. J.; Leverson, J. D.; Boghaert, E. R.; Ackler, S. L.; Catron, N. D.; Chen, J.; Dayton, B. D.; Ding, H.; Enschede, S. H.; Fairbrother, W. J.; Huang, D. C.; Hymowitz, S. G.; Jin, S.; Khaw, S. L.; Kovar, P. J.; Lam, L. T.; Lee, J.; Maecker, H. L.; Marsh, K. C.; Mason, K. D.; Mitten, M. J.; Nimmer, P. M.; Oleksijew, A.; Park, C. H.; Park, C. M.; Phillips, D. C.; Roberts, A. W.; Sampath, D.; Seymour, J. F.; Smith, M. L.; Sullivan, G. M.; Tahir, S. K.; Tse, C.; Wendt, M. D.; Xiao, Y.; Xue, J. C.; Zhang, H.; Humerickhouse, R. A.; Rosenberg, S. H.; Elmore, S. W. Abt-199, a Potent and Selective BCL-2 Inhibitor, Achieves Antitumor Activity While Sparing Platelets. *Nat. Med.* **2013**, *19*, 202−208.
(12) Reymond, J. L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722−730.
(13) 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.
(14) 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.
(15) 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.* **2001**, *46*, 3−26.
(16) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−1097.
(17) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.
(18) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K(I) Data. *J. Med. Chem.* **2012**, *55*, 5165−5173.
(19) Landrum, G. *RDKit: Open-Source Chemoinformatics*. http://www.rdkit.org (accessed Nov 29, 2018).
(20) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(21) Schmidt, D.; Bernat, V.; Brox, R.; Tschammer, N.; Kolb, P. Identifying Modulators of CXC Receptors 3 and 4 with Tailored Selectivity Using Multi-Target Docking. *ACS Chem. Biol.* **2015**, *10*, 715−724.
(22) Schmidt, D.; Gunera, J.; Baker, J. G.; Kolb, P. Similarity- and Substructure-Based Development of Beta2-Adrenergic Receptor Ligands Based on Unusual Scaffolds. *ACS Med. Chem. Lett.* **2017**, *8*, 481−485.
(23) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S. Probing the Bioactivity-Relevant Chemical Space of Robust Reactions and Common Molecular Building Blocks. *J. Chem. Inf. Model.* **2012**, *52*, 1167−1178.
(24) 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.
(25) Pevzner, Y. U.; Ihlenfeldt, W. D.; Nicklaus, M. Synthetically Accessible Virtual Inventory (SAVI). *Abstr. 253rd Am. Chem. Soc. Nat. Meet.* **2017**, 141.
(26) Humbeck, L.; Weigang, S.; Schafer, T.; Mutzel, P.; Koch, O. CHIPMUNK: A Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, Exploitable for Protein-Protein Interaction Modulators. *ChemMedChem* **2018**, *13*, 532−539.
(27) Rarey, M.; Stahl, M. Similarity Searching in Large Combinatorial Chemistry Spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497−520.
(28) Pottel, J.; Moitessier, N. Customizable Generation of Synthetically Accessible, Local Chemical Subspaces. *J. Chem. Inf. Model.* **2017**, *57*, 454−467.
(29) Batiste, L.; Unzue, A.; Dolbois, A.; Hassler, F.; Wang, X.; Deerain, N.; Zhu, J.; Spiliotopoulos, D.; Nevado, C.; Caflisch, A.

DOI: 10.1021/acs.jcim.8b00737
J. Chem. Inf. Model. 2019, 59, 644−651

Chemical Space Expansion of Bromodomain Ligands Guided by in Silico Virtual Couplings (Autocouple). *ACS Cent. Sci.* **2018**, *4*, 180−188.
(30) Nikitin, S.; Zaitseva, N.; Demina, O.; Solovieva, V.; Mazin, E.; Mikhalev, S.; Smolov, M.; Rubinov, A.; Vlasov, P.; Lepikhin, D.; Khachko, D.; Fokin, V.; Queen, C.; Zosimov, V. A Very Large Diversity Space of Synthetically Accessible Compounds for Use with Drug Design Programs. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 47−63.
(31) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and Searching 10(20) Synthetically Accessible Structures. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 341−350.
(32) Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, *49*, 270−279.
(33) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. Leap into the Pfizer Global Virtual Library (PGVL) Space: Creation of Readily Synthesizable Design Ideas Automatically. *Methods Mol. Biol.* **2011**, *685*, 253−276.
(34) Godfrey, A. G.; Masquelin, T.; Hemmerle, H. A Remote-Controlled Adaptive Medchem Lab: An Innovative Approach to Enable Drug Discovery in the 21st Century. *Drug Discovery Today* **2013**, *18*, 795−802.
(35) 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.
(36) Boehm, M.; Wu, T. Y.; Claussen, H.; Lemmen, C. Similarity Searching and Scaffold Hopping in Synthetically Accessible Combinatorial Chemistry Spaces. *J. Med. Chem.* **2008**, *51*, 2468−2480.
(37) Yu, N.; Bakken, G. A. Efficient Exploration of Large Combinatorial Chemistry Spaces by Monomer-Based Similarity Searching. *J. Chem. Inf. Model.* **2009**, *49*, 745−755.
(38) Chevillard, F.; Stotani, S.; Karawajczyk, A.; Stanimira, H.; Pardon, E.; Steyaert, J.; Tzalis, D.; Kolb, P. Interrogating Dense Ligand Chemical Space with a Forward-Synthetic Library, submitted for publication.
(39) TechStartups How Much Data Do We Create Every Day?. https://techstartups.com/2018/05/21/how-much-data-do-we-create-every-day-infographic/ (accessed Nov 24, 2018).
(40) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.
(41) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.
(42) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3*, 1237−1245.
(43) Schneider, G. Automating Drug Discovery. *Nat. Rev. Drug Discovery* **2017**, *17*, 97−113.
(44) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic Ai. *Nature* **2018**, *555*, 604−610.
(45) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23*, 5966−5971.
(46) 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.
(47) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations. *Chem. Sci.* **2019**, DOI: 10.1039/C8SC04175J.

DOI: 10.1021/acs.jcim.8b00737
J. Chem. Inf. Model. 2019, 59, 644−651