The next level in chemical space navigation: going far beyond enumerable compound libraries

Torsten Hoffmann[1] **and Marcus Gastreich**[2]

[1] Taros Chemicals GmbH & Co. KG, Emil-Figge-Str. 76a, 44227 Dortmund, Germany
[2] BioSolveIT GmbH, An der Ziegelei 79, 53757 Sankt Augustin, Germany

Recent innovations have brought pharmacophore-driven methods for navigating virtual chemical spaces, the size of which can reach into the billions of molecules, to the fingertips of every chemist. There has been a paradigm shift in the underlying computational chemistry that drives chemical space search applications, incorporating intelligent reaction knowledge into their core so that they can readily deliver commercially available molecules as nearest neighbor hits from within giant virtual spaces. These vast resources enable medicinal chemists to execute rapid scaffold-hopping experiments, rapid hit expansion, and structure–activity relationship (SAR) exploitation in largely intellectual property (IP)-free territory and at unparalleled low cost.

Introduction

For a long time, computational chemists have attempted to propose new ideas for lead molecules by tapping into novel IP spaces using techniques such as hit expansion, SAR exploitation, and scaffold hopping, yet the past 25 years have shown us that current computational algorithms and methods simply might not be capable of identifying highly innovative nearest neighbor molecules with much success. However, novel computational approaches now make it possible to identify new molecules that gain wide acceptance from medicinal chemists. Past experience has shown that it is decisive to incorporate synthetic knowledge into pharmacophore-based similarity searches in huge virtual chemistry spaces. This has been done using elegant computational algorithms that are extremely fast and easy to use, and that can take pharmacophore-based information into consideration. It has proved equally important to involve medicinal chemists during the generation of results *in silico*. The best currently available methods can search spaces close to the 4 billion molecule mark, and offer guaranteed delivery of successfully synthesized, tangible compounds, making rapid biological characterization from such enormous virtual chemistry spaces a realistic possibility [1].

Corresponding author. Gastreich, M. (gastreich@biosolveit.de)

Lead optimization projects in drug discovery often reach a dead end and result in failure. These projects not only try to exploit hitherto undiscovered modes of action, but may also examine the possibility of larger molecules as acting agents, such as peptides, macrocycles, biologics, and their conjugates. However, many research programs simply suffer from a lack of fast access to novel chemical classes that may display highly potent and selective pharmacological effects. On the one hand, access to larger molecular spaces is required to increase the likelihood of finding something 'interesting', whereas, on the other, methods for searching these spaces must be quicker, more efficient, and easy to use. Current chemical spaces range in size from a few thousand to 10^5 molecules offered by specialized suppliers offer ('small'), up to 10^8 ('large') for supplier pools, such as Molport, Chemspace, eMolecules, and others. Sizes considerably beyond 10^8 molecules ('giant') [2–4] can be considered as not practically tractable with traditional methods for various reasons that we discuss further below.

The pressure to be grand: novelty and IP

Beyond the obvious therapeutic focus, novelty is equally important. Only new IP can be patented and, thus, generate profit to finance further research in the search of new therapeutics. The solution to the search space size problem is to expand the realms of possibility using virtual molecules. One of the most prominent examples is the

https://doi.org/10.1016/j.drudis.2019.02.013

generic database (GDB) approach by Jean-Louis Reymond's group in Berne, Switzerland. His lab computationally generated all possible organic molecules under boundary constraints and applied various filters to avoid the creation of unwanted chemistry, such as multiple annealed small rings and instable element combinations. Particularly noteworthy are the graphical representations of chemical spaces in the respective publications (e.g., [5]).

Although rule-based generation (e.g., as conducted in [6] for macrolide scaffolds) is one approach to create an increasing number of molecular ideas, artificial intelligence (AI) methods have also recently gained attention for creating more novel IP in virtual molecular matter from a starting point [7]. Irrespective of the general approach to increasing the size of libraries by applying *in silico* techniques, researchers are now ready to create hundreds of millions (10^8) virtual compounds ([8,9] and references therein).

However, mere size causes several challenges for the way in which these compound numbers are processed or searched. Data-handling issues already arise during computer processing and in file or database storage; file sizes easily surpass the terabyte mark, searches, even in cloud-based environments, foreseeably take too much time: a 2012 publication by the Reymond group invested 100 000 central processing unit (CPU) hours on 360 processors in parallel to enumerate molecules with up to 17 heavy atoms [10], documenting the enormous efforts needed with such large compound pools. To give an example, a comparably small, 1-million compound subset of the GDB-13 [10] in uncompressed 3D SD format requires almost 1 GB hard disk space; Ruddigkeit *et al.* [11] stated that the GDB-17 molecules (1.66×10^{11} molecules) require ~400 GB disk space in zipped format. Thus, 10^{20} molecules would require $\sim 2 \times 10^8$ GB or 200 000 TB in compressed format; van Hilten *et al.* supplied more estimated storage sizes in their recent review [12]. In addition, such large numbers of enumerated virtual molecules have considerable effects on the time needed to search through the molecular contents of the space. State-of-the-art research, using computer code that is optimized for the hardware and incorporates numerous advanced speed-up strategies has shown that the timings for one substructure search in a few hundred million virtual compounds can be reduced to a few seconds [13]. However, considering the linear scaling behavior of this type of search, even these optimized approaches can no longer be used for the larger space sizes of 10^{12} or 10^{20} compounds [2] that are emerging now.

Experience has shown that the principle of enumeration can also cause patentability concerns because of the finite nature of classical compound collections: the limits of the collection define the limits of the IP content. To be awarded a patent, it is legally required to demonstrate novelty, utility, and nonobviousness. Therefore, using publically disclosed, finite, enumerated sets, such as the GDB databases, can reduce the chances of obtaining a patent. However, even using nonenumerated chemical spaces does not hold any guarantees because the process of patent awarding requires the patent counsel, and decisions are always made on a case-by-case basis [14].

Using combinatorics to go beyond enumeration: solving the synthesizability problem

Another concern with many *in silico* proposed compounds is the synthetic accessibility of the results. This can be entirely unclear and often prohibitive [9]. Machine-learning techniques are currently being used to help in this area: one recent method developed by Waller *et al.* [15] examines synthetic accessibility by retrosynthetic considerations and uses this information to train a proposal engine. The authors trained AI algorithms with millions of reactions from Elsevier's Reaxys database to predict promising synthetic pathways. The routes that the machine proposes were said to be 'on par with [previously] reported routes'. Yet, the IP-related concerns remain, and any algorithm that relies on enumeration (i.e., processing individual molecules; Fig. 1c) will suffer from prohibitively long runtimes for very large numbers of molecules.

One modern approach in the synthetic organic chemistry lab that avoids enumeration utilizes combinatorial expansion instead in the form of DNA-encoded libraries (DELs). The underlying technique for library creation can cover very large numbers of molecules [16,17]. Here, millions (or more) of combinatorially generated, putative small molecule binders are tagged with a 'signature' combination of DNA snippets, where the DNA acts like a barcode and, thus, is an unambiguous identifier. When screening against an immobilized protein target, those ligands that bind will stay bound and all others can be washed off. Subsequent PCR amplification and read-out of the barcodes of the binders enables scientists to identify active small molecules. Nevertheless, using computers as proposal engines remains of interest to save time and resources before investing into real wet lab time.

Combining fragments: a reborn approach?

Retrosynthetically motivated reassembly of fragments *in silico* has emerged as a method to overcome the computational challenges described earlier. The idea is as simple as it is striking, especially if one sees the analogy between virtual fragments in the computer and building blocks in the wet laboratory: taking only 1000 fragments that are allowed to combine via two reactions to form ('*de novo*') molecules, an impressive number of 1 billion ($1000 \times 1000 \times 1000$ or 10^9) virtual possibilities are already created. Adding more fragments and reactions cause an explosion in the size of the search space.

One of the first to pursue this idea was Xiao Lewell and coworkers with their popular RECAP approach [18]. After it was published, it was quickly implemented in various computer programs. The initial idea was to break down existing molecules by applying retrosynthetic shredding by the computer and then to reassemble novel compounds in a Lego-like fashion. Later on, more elaborate approaches that took existing knowledge about synthetic pathways into account were conceived [19–21], and the associated spaces made available to the general public. For example, in 2005, Nikitin *et al.* [22] published an approach that used catalog/vendor building blocks in a combinatorial *de novo* design computer program; the authors demonstrated their success by deriving micromolar HIV integrase inhibitors. However, because the synthetic tractability of results proposed by these methods remained an issue, the entire idea of exploiting combinatorics combinatorially was wrongfully condemned [23] and fell out of favor, until now, where the approach of combining fragments is experiencing a come back in this age of 'Big Data'.

The key to success: robust synthetic chemistry, knowledge assembled in the lab

Fragment-based approaches have regained new popularity in chemical space research and deliver compounds that are synthetically

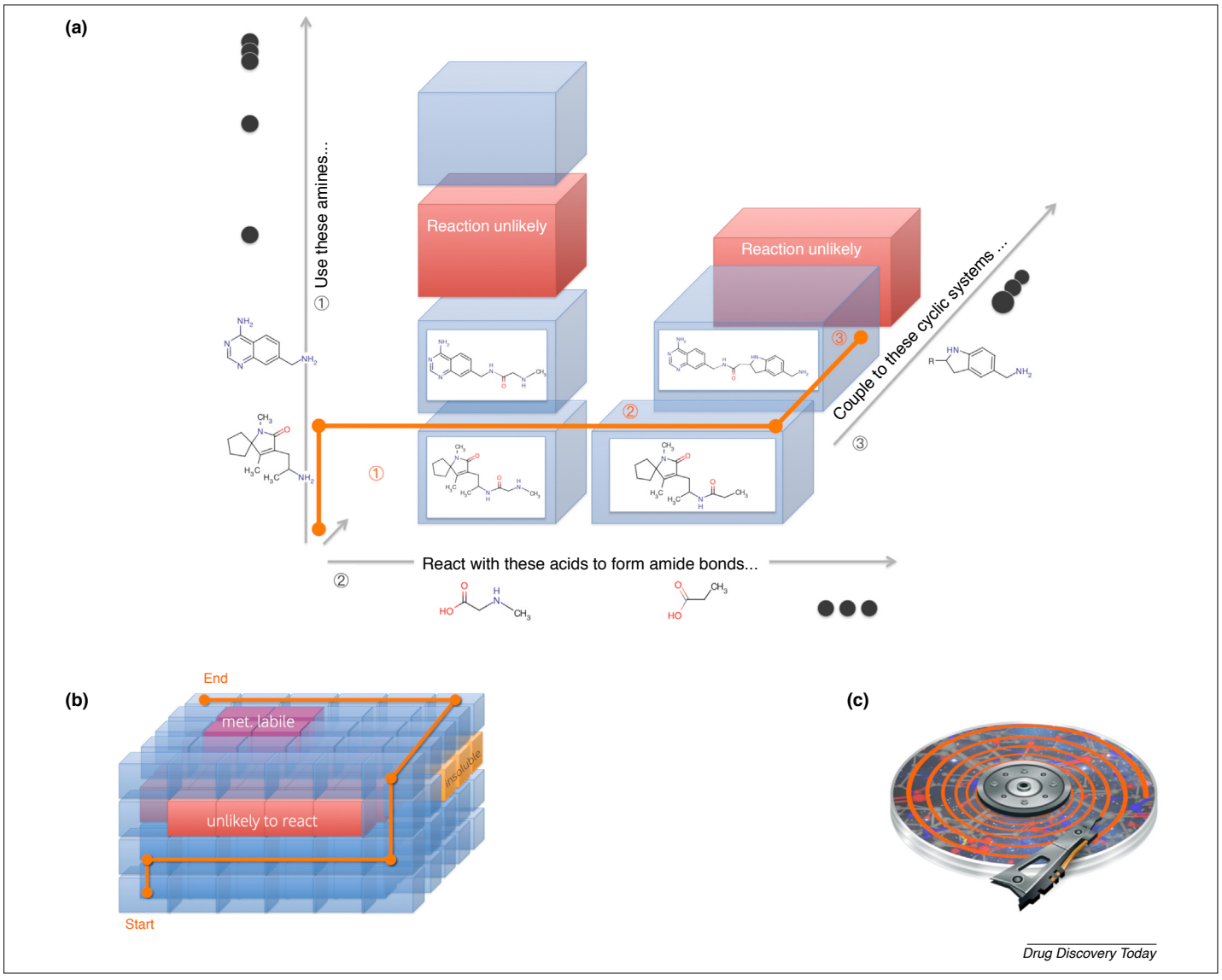

FIGURE 1

Combinatorial *vs.* enumerated chemical space navigation. **(a)** Navigation through reaction-driven combinatorial spaces (schematically): Starting from a 'seed' fragment (a building block, such as the spiro-compound in the lower left), novel compounds are created along a directed navigation path through a multidimensional hypercube. The (orange) path describes the sequence of formed virtual reactions, here an amide forming followed by a ring coupling; the dark circles denote more instances of the building blocks. Thus, resulting molecules (on boxes) are built from connecting virtual building blocks ('fragments') and searched at the same time. Enumeration can be omitted (i.e., it becomes unnecessary to 'visit' every virtual compound that can in principle be built) and, therefore, larger virtual spaces can be created and mined in less time. Computational strategies typically do not operate on the actual chemical structures but on so-called descriptors that encode physicochemical properties. This principle idea is, for example, implemented in the REAL Space Navigator using Feature Trees [36,37], or reaction driven in the DOGS approach [20]. **(b)** Extrapolation from (a). Creation while navigating through a combinatorial chemical space along a path (orange): certain regions in this search space are not visited because the necessary reactions are very unlikely to occur; in addition, estimated or properties that could be computed in look-ahead strategies are unwanted (e.g., a predicted surpassing of requested molecular weight, metabolic instability, an impossibility to comply with a pharmacophore, or unlikely solubility when adding certain building blocks). **(c)** Schematic display of storage and searches in enumerated chemical spaces that are physically kept on a hard disk in the computer: in contrast to combinatorial space navigation as in (a) and (b), searching large numbers of stored molecules involves individual processing of every molecule, for example along the (orange) path. The storage capacity is limited by both handling times and storage spaces. As stated in [11], the generation of the GDB-17 database [10] required 400 GB of storage capacity in compressed SMILES format of molecules. Today's limit for processing enumerated numbers of compounds is $\sim 10^9$ molecules.

accessible by design. The key to rendering the combinatorial approach useful is as simple as it is elegant: Instead of looking backwards, applying generalized schemes of historic chemistry knowledge to the assembly of novel IP, an increasing number of organizations are looking forward. These approaches extrapolate from robust and well-tracked in-house synthetic chemistry knowledge towards creating those virtual molecules that are tangible with high likelihood of synthetic access in the lab (see Fig. 2 for a comparison of large chemical spaces and their evolution over time and Table 1 for more details including references). The 'paradigm shift' alluded to earlier lies in acknowledging the necessity for synthetic accessibility and implementing it as an integral part of the computer algorithms that navigate

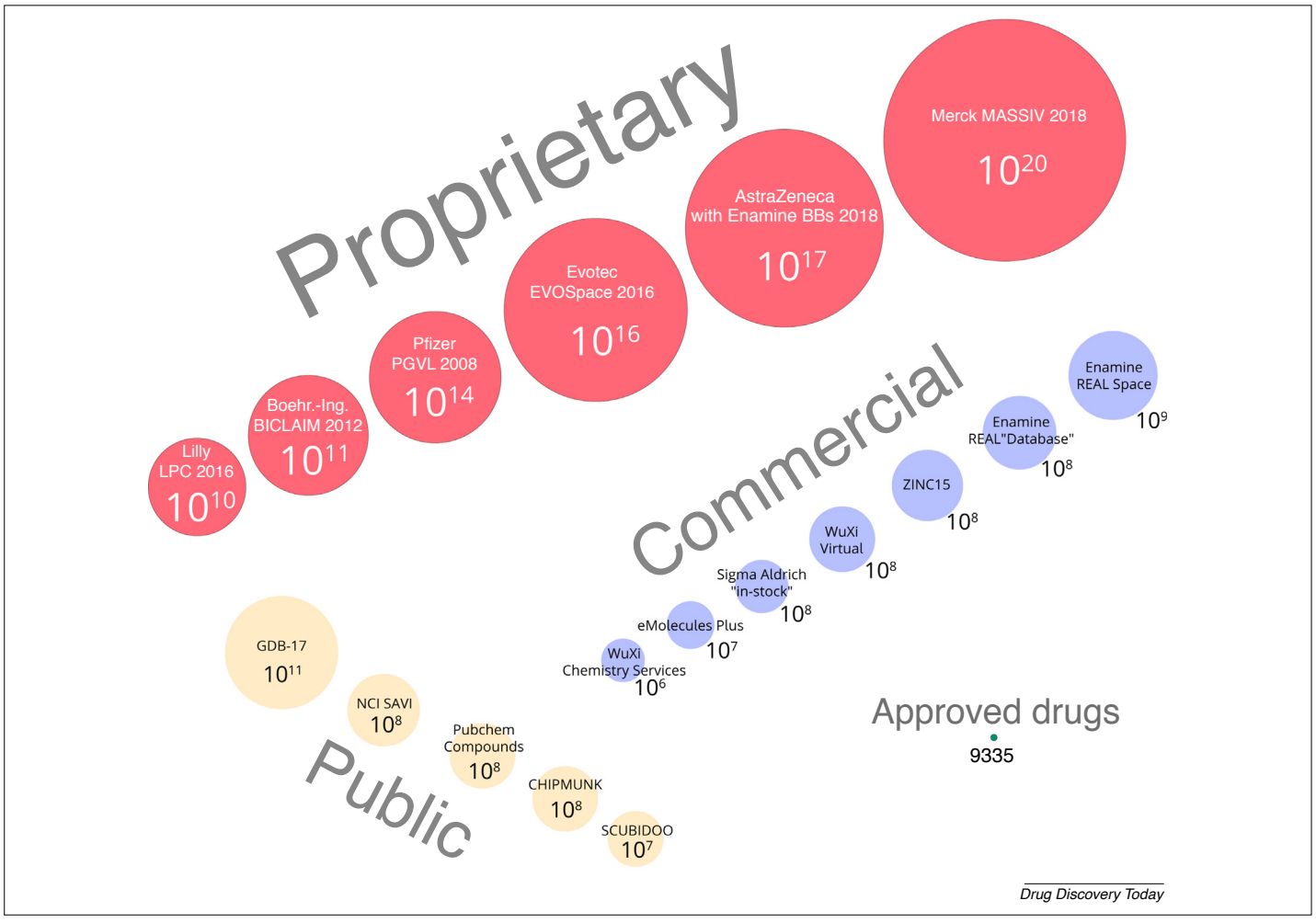

FIGURE 2

Classification and (rounded) orders of magnitude of sizes for selected chemical spaces compared with the number of approved drugs (lower right) in DrugBank on November 13, 2018 [40]. The radii of the circles reflect logarithmic representations of compound numbers. To clarify the fast historic progress, the numbers within or next to the circles denote the order of magnitude of the space contents at the time of the respective first available publication (it is likely that the spaces have grown further over time; e.g., Pfizer's PGVL included 10^{18} compounds by 2012 [41]). The displayed 'academic' spaces have been entirely, or in parts, made available to the general public. For more information, origin details, references, and abbreviations, see Table 2 in the main text. Abbreviations: BBs, building blocks; BICLAIM, Boehringer Ingelheim Comprehensive Library of Accessible Innovative Molecules; Boehr.-Ing., Boehringer-Ingelheim; CHIPMUNK, CHemically feasible In silico Public Molecular UNiverse Knowledge base; GDB-17, Generic Database (of molecules with up to 17 heavy atoms); LPC, Lilly Proximal Collection; MASSIV, Merck AcceSSIble InVentory; PVGL, Pfizer Global Virtual Library; SAVI, Synthetically Accessible Virtual Inventory; SCUBIDOO, Screenable Chemical Universe Based on Intuitive Data OrganizatiOn.

through a multidimensional hypercube (Fig. 1a,b). What used to be completely decoupled from one another, namely the search descriptor and compound synthesis, are now closely interlinked and, as a result, the approach has now gained broad acceptance: The general idea was first published by a team at Pfizer [24], who created a reaction-driven chemical space of 10^{16} virtual compounds ('PGVL'); they documented several examples of active compounds that they were successfully able to mine from it. This first publication was quickly followed by a similar approach from Boehringer-Ingelheim [4,25], who reported on active compound generation with novel IP; the hits found were rapidly optimized into lead compounds against GPR119, a challenging novel GPCR target involved in insulin secretion. The largest in-house chemical search space to date was recently been presented by Merck KGaA during the celebration conference on its 350th anniversary [2,3]. The so-called MASSIV space, if one could ever enumerate it, offers a striking 10^{20} molecular possibilities, and the authors reported on the successful design of active compounds created using the encoded corporate in-house synthetic procedures. It could reassuringly be demonstrated that the results were synthetically accessible and the compounds were delivered in vials within a short period of time. As an example, one fragment hit was rapidly transformed into a single digit micromolar compound for one hitherto undisclosed target; a respective publication is currently being prepared [26]. It must be stressed that the robustness of the reactions and/or transformations is a pivotal element in this procedure, and that the quality and diversity and/or coverage of such chemical spaces will increase with the creativity and novelty of the chemistry behind them (see [27,28] for recent overviews on chemical reactions in medicinal chemistry). Understandably, only a few organizations publish details on the reactions that they work with, and, what is even more relevant, the yields and potential variations in experimental conditions. Therefore, the ultimate measure of quality for a 'large'/'giant' chemical

TABLE 1

Examples of large virtual and tangible/accessible chemistry spaces[a]

Organization	Name	Number of compounds	Hits purchasable from creator[b]	Refs
Merck KGaA	MASSIV	10^{20}		[2]
AstraZeneca	AZ Space with Enamine BBs[c]	10^{17}		[44]
Evotec	EVOspace	1.6×10^{16}	(Yes)[d]	[45]
Pfizer	PGVL	3×10^{12}		[24,41]
Boehringer-Ingelheim	BICLAIM	5×10^{11}		[4,25]
Lilly	Lilly LPC	2×10^{11}		[46]
University of Berne	GDB-17	2×10^{11}		[10]
NCI	SAVI	3×10^{8}		[47]
UCSF	ZINC15	2×10^{8}		[48]
University of Dortmund	CHIPMUNK	1×10^{8}		[49]
Enamine	REAL 'Space'	4×10^{9}	Yes	[30]
University of Marburg	SCUBIDOO	2.1×10^{7}		[50]
Enamine	REAL 'Database'	6.8×10^{8}	Yes	[1]
WuXi Apptec	WuXi Virtual	1×10^{8}	Yes	[51]
NCBI	PubChem Compounds	9.6×10^{7}		[52]
Aldrich Market Select	Sigma Aldrich 'in-stock'	1.4×10^{7}	Yes	[53]
eMolecules	eMolecules Plus	5.9×10^{6}	Yes	[54]
WuXi	WuXi Chemistry Services	3×10^{6}	Yes	[55]

[a] The number of molecules in large chemical spaces today ranges from a few million that are commercially available or have proven to be accessible, across billions of tangible/purchasable compounds to huge, largely undisclosed in-house corporate spaces. The selection of chemical spaces is based on publicly available information.
[b] Those spaces from which compounds can be obtained commercially are annotated with 'yes', all others imply a 'no'.
[c] Abbreviation: BBs, building blocks.
[d] EVOspace searches are a CRO service offered by Evotec.

space is how much of the relevant chemical space it covers (which is related to its space size) and the percentage of compounds that can be synthesized from it. Affinity is both query and target dependent and not an inherent property of the space, unless it is a small or focused space that has been created for a certain target class, for example.

Returning to the practical aspects of obtaining novel molecules from chemical spaces, it is obvious how these methods can work with in-house data. However, how much more desirable would it be to be able to simply purchase molecules on demand from commercial suppliers, who, in fact, operate on vast virtual chemical spaces based on established synthetic procedures? This recently became possible (see Fig. 3 for an overview of current chemical spaces that comprise purchasable or highly tangible compounds). The Ukrainian company Enamine, a compound provider and contract research organization, recently perfected this approach. Its REAL Space concept [1] (not to be confused with the much smaller REAL *Database*) uses existing building blocks and reactions taken directly from their synthetic procedures for which they have quality control records that reach back several years. The creation of such a space is taken care of using a software named CoLibri [29]. Using a conservative quality threshold, the company guarantees a delivery success rate of >80% within approximately 3 weeks. The combinatorial nature of their chemistry creates an enormous space that is largely accessible in reality. Researchers can order directly from the supplier and, upon ordering, obtain ownership of the associated IP. If enumerated, the REAL Space would comprise more than 3.8 billion compounds. Even at 80%, this still translates to more than 3 billion compounds being deliverable in theory. An associated chemical space search software from Germany's BioSolveIT, the REAL Space Navigator [30], can be downloaded for free and can navigate through these billions of possibilities within 2–3 min, delivering 1000 compounds per run.

A priori, the mere size of a chemical space does not guarantee relevance for a specific target or a certain amount of diversity; not even usefulness is ensured. However, probing the quality of the REAL Space by conducting a brief building-block analysis and applying the boundary condition that only one or two-step synthetic routes are used, molecules displaying typical ranges for drug-like properties emerge (Fig. 4). Furthermore, to obtain relevant hit molecules, pharmacophore-like features (see later) can be used as guidance during navigation of the REAL Space. Result molecules are generated together with their similarity score with respect to the query. Finally, subsequent filtering for PAINS and aggregators [31–33], or further visual and computational assessment can be conducted easily on thousands of resulting molecules using traditional methods. However, before selected compounds are ordered, quality control and expert analyses of results should be carried out by medicinal chemists [34,35].

The ultimate problem to address is not necessarily to create a virtual space that covers all possibilities, but rather to find in, or to mine from, a virtual space neighboring compounds that are most relevant based on pharmacophores. The computational challenge lies in navigating chemical spaces efficiently and delivering attractive, molecular proposals that are relevant for a research project. Algorithms are needed that go beyond traditional substructure searches. A few, extremely fast computer programs are now available (see Table 2 for an overview of some of the fastest available methods). The approaches of both Merck and Enamine use modern variants of fuzzy pharmacophore-like descriptors, so-called 'Feature Trees' [36]. This strategy, an idea developed by Matthias Rarey, which later further evolved in a collaboration with Roche [37], uses a tree-based description of molecules, in combination with so-called 'dynamic programming techniques'. The latter are well known from protein alignment strategies, are extremely fast, and, by design, will find the best possible alignment, a measure for 'spatial' molecular similarity.

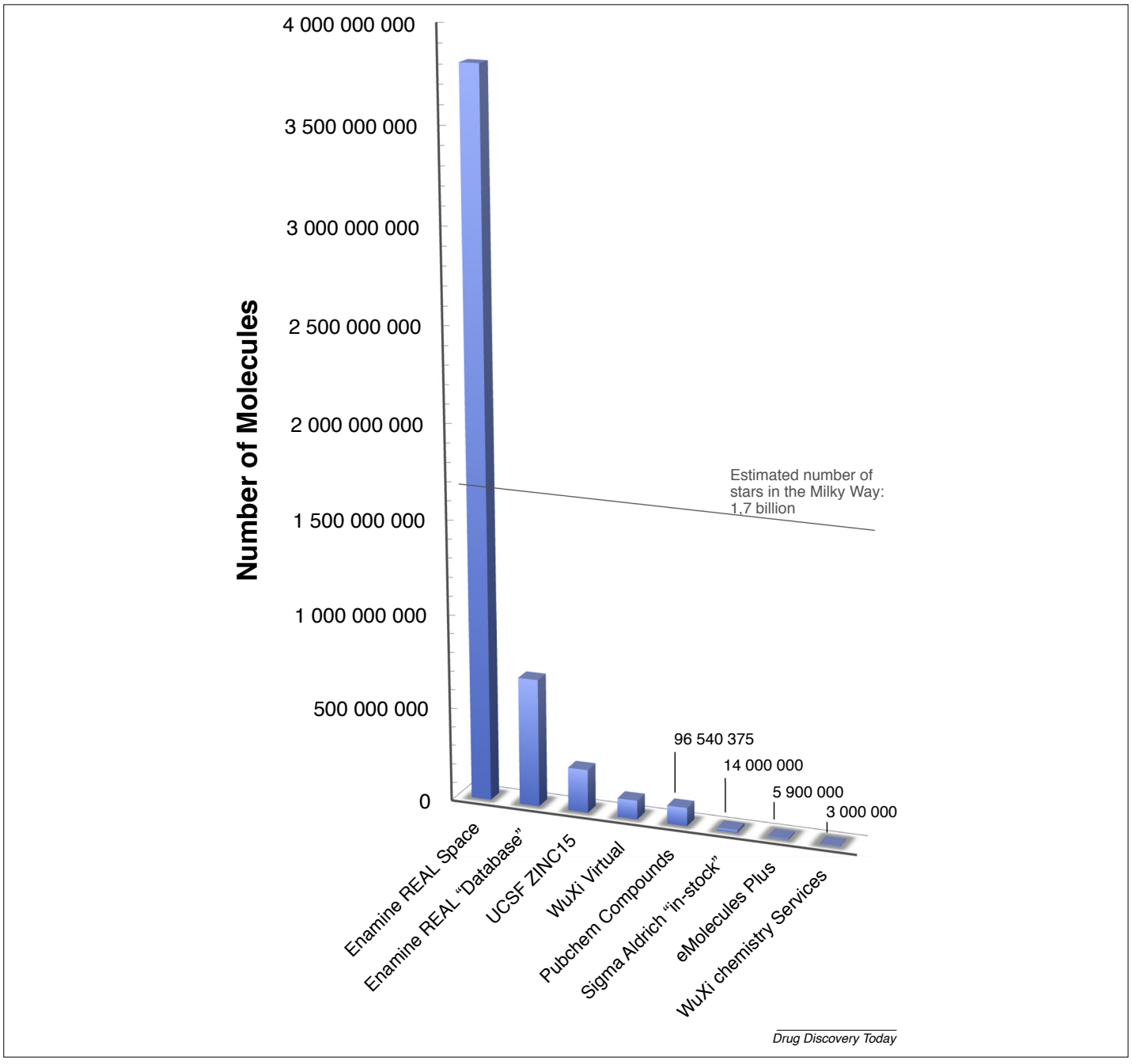

FIGURE 3

Illustration of chemical spaces of commercially available or tangible compounds. Compounds from highly accessible chemical spaces exceed the PubChem collection of compounds by orders of magnitude. The REAL Space concept is a hybrid method, with 80% accessible compounds from a virtual space and covers approximately the doubled size of the number of estimated stars in the Milky Way [42]. The size of the entire relevant chemical universe is estimated to be 10^{63} compounds [43].

Wherein lie the limitations of the approach? Navigation within the space requires an *in silico* description of what a molecule is, a so-called 'descriptor'. Search results will be subject to the limitations of a descriptor. In the case of Feature Trees [36], it is the inability to distinguish ring substitution patterns and stereochemistry. Therefore, several users have reported (e.g., [4]) orthogonal post-processing of results with 3D methods, such as 3D superposition, for enhancement. Another current drawback with large, combinatorial chemical spaces is their current inaccessibility with respect to full substructure searches and comprehensive analyses of typically relevant parameters. Questions such as: 'What is the distribution of logP, or the number of acceptors/donors, across the whole space?' cannot be answered. However, this also holds true for large enumerated spaces. Lucas *et al.* [38] conducted a careful analysis on the purchasable chemical space that, at that time, comprised ~68 million (6.8×10^7) molecules. However, several current, combinatorially generated spaces surpass the 10^7-molecule mark by a factor of ten or more orders of magnitude (Fig. 2)

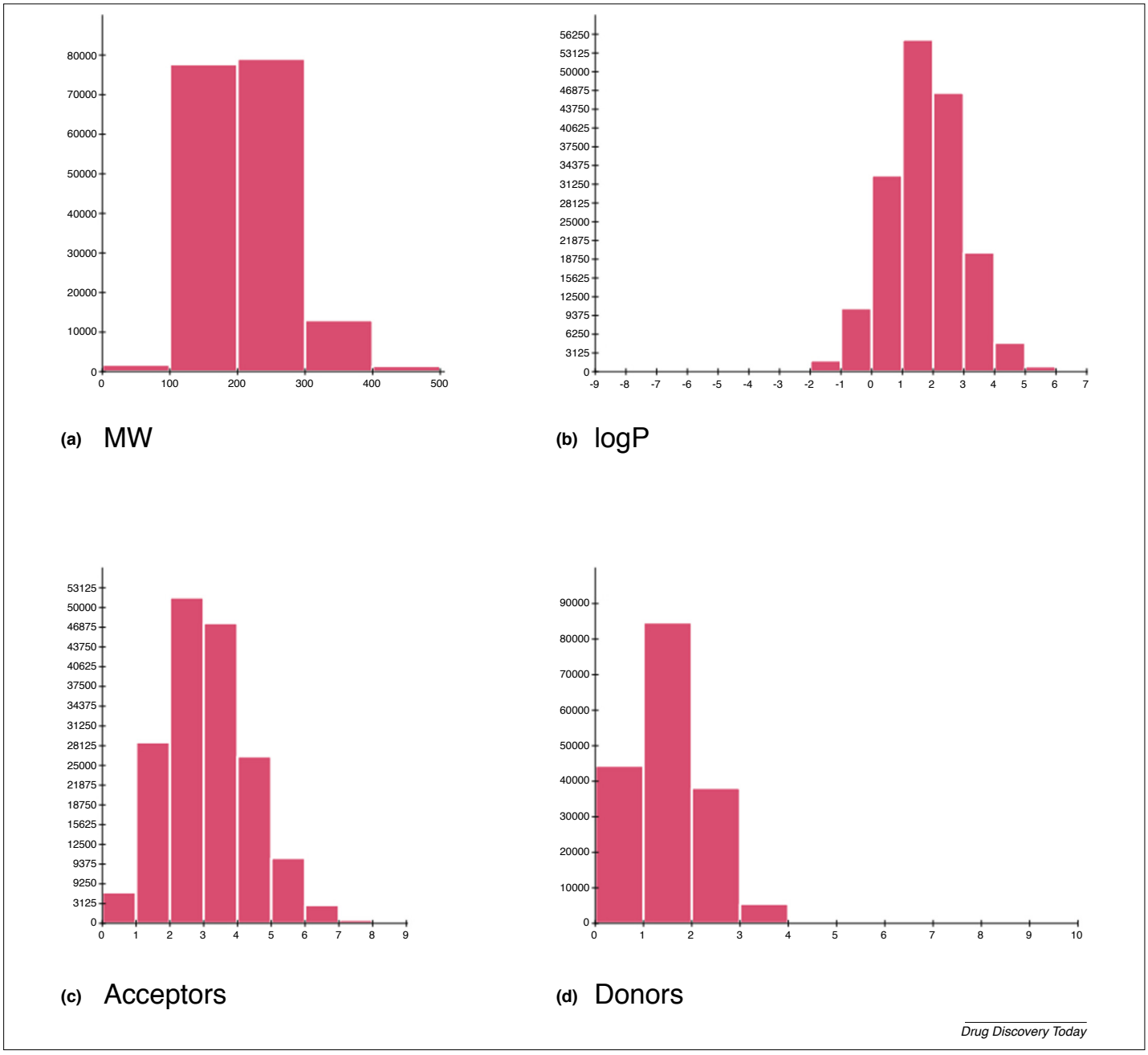

FIGURE 4

Distribution of physicochemical properties of the building blocks in the REAL Space [1]. (a) The molecular weight (MW) ranges from small fragment-like building blocks to 500 Da and averages at ~250 Da; with one or two-step reactions, drug-like molecular weights are reached. (b) The logP values show a peak slightly below 2, backing the assumption of relevance in small-molecule research; the distribution of the acceptor (c) and donor (d) counts range between 0–7 and 0–4, respectively.

TABLE 2

Examples of search technologies for chemistry spaces[a]

Tool name	Supplier	Rationale behind	Refs
DOGS/CATS	Schneider Group	Reaction-driven space generation/fuzzy 2D biophore similarities	[20,56]
MadFast	ChemAxon	2D similarity	[57]
Arthor	NextMove Software	Hardware-optimized similarity/substructure	[13]
FastROCS	OpenEye Scientific	Gaussian descriptions of 3D shape and property alignment of molecules on GPUs[b]	[58,59]
REAL Space Navigator (FTrees/-FS & CoLibri)	BioSolveIT	Validated reactions plus building blocks and combinatorics, fuzzy pharmacophores	[29,30,37]

[a] The selection of software tools is based on publicly available information.
[b] Abbreviation: GPU, graphics processing unit.

TABLE 3

Early feedback after the REAL Space Navigator launch in March 2018 from anonymous pharmaceutical companies and Enamine[a]

	Big Pharma 1	Big Pharma 2	Big Pharma 3	Enamine Total
Number of requested compounds	20–80	$100 < n < 1000$	n/a	2016
Average number of compounds ordered	n/a	n/a	n/a	25
Average number of days from order to delivery	n/a	n/a	n/a	27
Average success rate	84%	>90%	91%	87%

[a] The reported figures refer to the early timeframe between March 2018, when the REAL Space Navigator was first launched, and November 2018 (~8 months). The average success rate (ratio between ordered versus delivered molecules, in percent) consistently lies >80%. Not all companies that were contacted agreed to disclose all details, and several did not supply figures at all; others were planning to order, or they did not record such statistics. The values reported by Enamine refer to March to September 2018. Several pharma companies verbally confirmed the delivery times to the authors.

and, therefore, remain inaccessible to such (or similar [39]) analysis. As an alternative it is possible, for example, to compute a few thousand results using diverse drugs or drug-like query compounds and analyze the results to arrive at a more quantitative assessment of the properties of a space. Filtering giant spaces for pK_a ranges, metabolic instability, and comparable parameters currently remains a postprocessing task on the results of the search before making the final compound order.

Purchasing the compounds from compound suppliers can be realized in a user-friendly way, similar to Google's image search: synthetic chemists can copy and paste a 2D molecular drawing into the REAL Space Navigator, run the search securely behind their corporate firewall on their own desktop, and assess the results immediately, avoiding delays from processes that involve multiple groups and departments that often take place sequentially. Table 3 lists some early statistics that were reported by both pharma companies and Enamine. To date, the REAL Space concept is the world leader in terms of readily purchasable molecules, both with respect to the number of molecules that can be ordered as well as computing times, with no preprocessing necessary.

Concluding remarks and outlook

We are convinced that the described approach has the potential to lead to a paradigm shift in drug discovery and molecular design of small-molecule therapeutics if it were applied extensively throughout both industry and academia. The successful combination of pharmacophore-based similarity searches in huge virtual chemistry spaces and the incorporation of synthetic knowledge in elegant computational algorithms enables medicinal chemists to execute rapid hit expansion, powerful exploitation of SAR knowledge, and innovative scaffold hopping. The approach is a demonstration of a long-awaited step change, documenting how the scientific application of medicinal chemistry and molecular design is now maturing from a traditional science into a more agile core discipline of drug discovery. This is just the beginning of a new era for drug discovery, because large virtual, purchasable, and tangible molecular compound spaces will continue to grow in size. The first iteration of the REAL Space exhibited 640 million compounds. Within less than 6 months, the space grew up to 3.8 billion molecules: the future is closer than we think.

Acknowledgments

We are grateful to Sally Hindle and Christian Lemmen for stimulating discussions and valuable input to this publication.

Declaration of interest

M.G. is an employee of BioSolveIT, manufacturer of one of the software approaches covered in this article. T.H. is a scientific advisor to BioSolveIT and Elsevier; both companies are mentioned in this article.

References

1 Enamine Enamine REAL Space and REAL Database. https://enamine.net/index.php?option5com_content&task5view&id5254 [Accessed 26 February 2019]
2 Krier, M. and Klingler, F.-M. (2018) *10^{20} Molecules — A Gigantic Pool of Possibilities at your Fingertips*. Curious Conference
3 Knehans, T. *et al.* (2017) *Abstract: Merck AcceSSible InVentory (MASSIV): In silico Synthesis Guided by Chemical Transforms Obtained through Bootstrapping Reaction Databases*. American Chemical Society
4 Wellenzohn, B. *et al.* (2012) Identification of new potent GPR119 agonists by combining virtual screening and combinatorial chemistry. *J. Med. Chem.* 55, 11031–11041
5 Awale, M. *et al.* (2013) MQN-mapplet: visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inf. Model.* 53 (2), 509–518
6 Zin, P.P.K. *et al.* (2018) Cheminformatics-based enumeration and analysis of large libraries of macrolide scaffolds. *J. Cheminform.* 10, 1–20
7 Segler, M.H.S. *et al.* (2018) Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* 4 (1), 120–131
8 Mullard, A. (2017) The drug-maker's guide to the galaxy. *Nature* 549, 445–447
9 Walters, W.P. (2019) Virtual chemical libraries. *J. Med. Chem.* 62 (3), 1116–1124
10 Ruddigkeit, L. *et al.* (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52, 2864–2875
11 Ruddigkeit, L. *et al.* (2013) Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* 53 (1), 56–65
12 van Hilten, N. *et al.* (2019) Virtual compound libraries in computer-assisted drug discovery. *J. Chem. Inf. Model.* 59 (2), 644–651
13 Sayle, R. *et al.* (2018) *Recent Advances in Chemical & Biological Seach Systems: Evolution vs. Revolution*. NetMove Softward
14 United States Patent and Trademark Office . *Patent Process Overview*, USPTO 2019; https://www.uspto.gov/patents-getting-started/patent-process-overview [Accessed 01 March 2019]
15 Segler, M.H.S. *et al.* (2018) Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610
16 Brenner, S. and Lerner, R.A. (1992) Encoded combinatorial chemistry (chemical repertoire/encoded libraries/commaless code). *Proc. Natl. Acad. Sci. U. S. A.* 89, 5381–5383
17 Favalli, N. *et al.* (2018) DNA-encoded chemical libraries –achievements and remaining challenges. *FEBS Lett.* 592, 2168–2180
18 Lewell, X.Q. *et al.* (1998) 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.* 38, 511–522
19 Degen, J. *et al.* (2008) On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* 3, 1503–1507

20 Hartenfeller, M. *et al.* (2012) DOGS: reaction-driven de novo design of bioactive compounds. *PLoS Comput. Biol.* 8, 1–12
21 Degen, J. and Rarey, M. (2006) FlexNovo: Structure-based searching in large fragment spaces. *ChemMedChem* 1 (8), 854–868
22 Nikitin, S. *et al.* (2005) A very large diversity space of synthetically accessible compounds for use with drug design programs. *J. Comput. Aided Mol. Des.* 19, 47–63
23 Kodadek, T. (2011) The rise, fall and reinvention of combinatorial chemistry. *Chem. Commun.* 47, 9757–9763
24 Boehm, M. *et al.* (2008) Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *J. Med. Chem.* 51, 2468–2480
25 Lessel, U. *et al.* (2009) Searching fragment spaces with feature trees. *J. Chem. Inf. Model.* 49, 270–279
26 Krier M. Personal communication 2019
27 Boström, J. *et al.* (2018) Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discov.* 17, 709–727
28 Blakemore, D.C. *et al.* (2018) Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* 10, 383–394
29 Lilienthal, M. *et al.* CoLibri 2018. www.biosolveit.de/CoLibri [Accessed 26 February 2019]
30 Enamine, BioSolveIT. Enamine REAL Space Navigator v2.1. https://www.biosolveit.de/realspacenavigator [Accessed 26 February 2019]
31 Blevitt, J.M. *et al.* (2017) Structural basis of small-molecule aggregate induced inhibition of a protein-protein interaction. *J. Med. Chem.* 60 (8), 3511–3517
32 Pouliot, M. and Jeanmart, S. (2016) Pan assay interference compounds (PAINS) and other promiscuous compounds in antifungal research. *J. Med. Chem.* 59, 497–503
33 Baell, J.B. and Holloway, G.A. (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740
34 Rognan, D. (2013) Towards the next generation of computational chemogenomics tools. *Mol. Inform.* 32, 1029–1034
35 Medina-Franco, J.L. *et al.* (2014) Balancing novelty with confined chemical space in modern drug discovery. *Expert Opin. Drug Discov.* 9, 151–165
36 Rarey, M. and Dixon, J.S. (1998) Feature trees: a new molecular similarity measure based on tree matching. *J. Comput. Aided Mol. Des.* 12, 471–490
37 Rarey, M. and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aided Mol. Des.* 15, 497–520
38 Lucas, X. *et al.* (2015) The purchasable chemical space: a detailed picture. *J. Chem. Inf. Model.* 55, 915–924
39 Saldívar-González, F.I. *et al.* (2019) Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. *J. Chem. Inf. Model.* 59 (1), 54–85
40 Wishart, D.S. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082
41 Hu, Q. *et al.* (2012) Pfizer global virtual library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb. Sci.* 14, 579–589
42 European Space Agency . *Second Data Release from ESA's GAIA Mission*, 2018; https://www.esa.int/Our_Activities/Space_Science/Gaia/Call_for_media_Second_data_release_from_ESA_s_Gaia_mission [Accessed 01 March 2019]
43 Bohacek, R.S. *et al.* (1996) The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* 16, 3–50
44 Grebner, C. (2018) Webinar: Exploration and Mining of Large Virtual Chemical Spaces. https://youtu.be/fMrI11SXwpU [Accessed 26 February 2019]
45 M.P. Mazanetz , et al. *KNIME-ing through the EVOspace of FTrees*, BioSolveIT/Evotec 2016; https://www.evotec.com/en/execute/science-pool/scientific-publications [Accessed 01 March 2019]
46 Nicolaou, C.A. *et al.* (2016) The Proximal Lilly Collection: mapping, exploring and exploiting feasible chemical space. *J. Chem. Inf. Model.* 56, 1253–1266
47 Ott, M. *et al.* (2018) Synthetically Accessible Virtual Inventory (SAVI) Database. https://cactus.nci.nih.gov/download/savi_download/ [Accessed 26 February 2019]
48 Sterling, T. and Irwin, J.J. (2015) ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* 55, 2324–2337
49 Humbeck, L. *et al.* (2018) CHIPMUNK: a virtual synthesizable small-molecule library for medicinal chemistry, exploitable for protein–protein interaction modulators. *ChemMedChem* 13, 532–539
50 Chevillard, F. and Kolb, P. (2015) 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.* 55, 1824–1835
51 WuXi Apptec . *WuXi AppTec Launches Advanced Library Compound Initiative On LabNetwork.com*, 2018; http://www.wuxiapptec.com/press/detail/357/18.html [Accessed 26 February 2019]
52 National Center for Biotechnology Information. PubChem. https://pubchemdocs.ncbi.nlm.nih.gov/about [Accessed 26 February 2019]
53 Merck. Aldrich Market Select Chemistry Services. 2019. www.sigmaaldrich.com/chemistry/chemistry-services/aldrich-market-select.html [Accessed 26 February 2019]
54 eMolecules Inc. eMolecules database download. www.emolecules.com/info/plus/download-database [Accessed 26 February 2019]
55 WuXi Apptec. WuXi Apptec Synthetic Chemistry. www.wuxiapptec.com/pha_disc_synthmedchem.html [Accessed 26 February 2019]
56 Schneider, G. *et al.* (1999) 'Scaffold-hopping' by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed. Engl.* 38, 2894–2896
57 ChemAxon. MadFast Similarity Search. https://chemaxon.com/products/madfast [Accessed 26 February 2019]
58 OpenEye Scientific FastROCS. www.eyesopen.com/molecular-modeling-fastrocs [Accessed 26 February 2019]
59 Rush, T.S. *et al.* (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* 48, 1489–1495