Computational approaches streamlining drug discovery

https://doi.org/10.1038/s41586-023-05905-z

Received: 21 April 2022

Accepted: 1 March 2023

Published online: 26 April 2023

Check for updates

Anastasiia V. Sadybekov[1,2] & Vsevolod Katritch[1,2,3] ✉

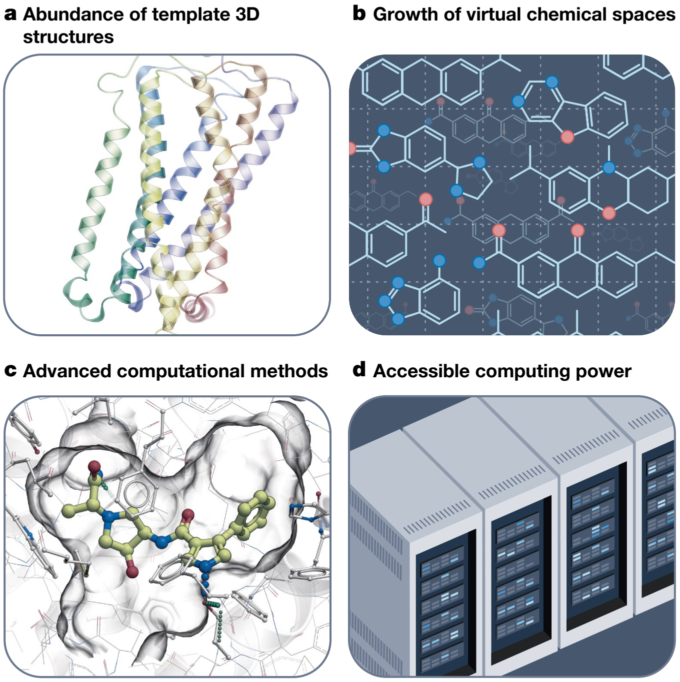
Computer-aided drug discovery has been around for decades, although the past few years have seen a tectonic shift towards embracing computational technologies in both academia and pharma. This shift is largely defined by the flood of data on ligand properties and binding to therapeutic targets and their 3D structures, abundant computing capacities and the advent of on-demand virtual libraries of drug-like small molecules in their billions. Taking full advantage of these resources requires fast computational methods for effective ligand screening. This includes structure-based virtual screening of gigascale chemical spaces, further facilitated by fast iterative screening approaches. Highly synergistic are developments in deep learning predictions of ligand properties and target activities in lieu of receptor structure. Here we review recent advances in ligand discovery technologies, their potential for reshaping the whole process of drug discovery and development, as well as the challenges they encounter. We also discuss how the rapid identification of highly diverse, potent, target-selective and drug-like ligands to protein targets can democratize the drug discovery process, presenting new opportunities for the cost-effective development of safer and more effective small-molecule treatments.

Despite amazing progress in basic life sciences and biotechnology, drug discovery and development (DDD) remain slow and expensive, taking on average approximately 15 years and approximately US$2 billion to make a small-molecule drug[1]. Although it is accepted that clinical studies are the priciest part of the development of each drug, most time-saving and cost-saving opportunities reside in the earlier discovery and preclinical stages. Preclinical efforts themselves account for more than 43% of expenses in pharma, in addition to major public funding[1], driven by the high attrition rate at every step from target selection to hit identification and lead optimization to the selection of clinical candidates. Moreover, the high failure rate in clinical trials (currently 90%)[2] is largely explained by issues rooted in early discovery such as inadequate target validation or suboptimal ligand properties. Finding fast and accessible ways to discover more diverse pools of higher-quality chemical probes, hits and leads with optimal absorption, distribution, metabolism, excretion and toxicology (ADMET) and pharmacokinetics (PK) profiles at the early stages of DDD would improve outcomes in preclinical and clinical studies and facilitate more effective, accessible and safer drugs.

The concept of computer-aided drug discovery[3] was developed in the 1970s and popularized by *Fortune* magazine in 1981, and has since been through several cycles of hype and disillusionment[4]. There have been success stories along the way[5] and, in general, computer-assisted approaches have become an integral, yet modest, part of the drug discovery process[6,7]. In the past few years, however, several scientific and technological breakthroughs resulted in a tectonic shift towards embracing computational approaches as a key driving force for drug discovery in both academia and industry. Pharmaceutical and biotech companies are expanding their computational drug discovery efforts or hiring their first computational chemists. Numerous new and established drug discovery companies have raised billions in the past few years with business models that heavily rely on a combination of advanced physics-based molecular modelling with deep learning (DL) and artificial intelligence (AI)[8]. Although it is too early yet to expect approved drugs from the most recent computationally driven discovery efforts, they are producing a growing number of clinical candidates, with some campaigns specifically claiming target-to-lead times as low as 1–2 months[9,10], or target-to-clinic time under 1 year[11]. Are these the signs of a major shift in the role that computational approaches have in drug discovery or just another round of the hype cycle?

Let us look at the key factors defining the recent changes (Fig. 1). First, the structural revolution—from automation in crystallography[12] to microcrystallography[13,14] and most recently cryo-electron microscopy technology[15,16]—has made it possible to reveal 3D structures for the majority of clinically relevant targets, often in a state or molecular complex relevant to its biological function. Especially impressive has been the recent structural turnaround for G protein-coupled receptors (GPCRs)[17] and other membrane proteins that mediate the action of more than 50% of drugs[18], providing 3D templates for ligand screening and lead optimization. The second factor is a rapid and marked expansion of drug-like chemical space, easily accessible for hit and lead discovery. Just a few years ago, this space was limited to several million on-shelf compounds from vendors and in-house screening libraries in pharma. Now, screening can be done with ultra-large virtual libraries

[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA. [2]Center for New Technologies in Drug Discovery and Development, Bridge Institute, Michelson Center for Convergent Biosciences, University of Southern California, Los Angeles, CA, USA. [3]Department of Chemistry, University of Southern California, Los Angeles, CA, USA. ✉e-mail: katritch@usc.edu



a Abundance of template 3D structures

b Growth of virtual chemical spaces

c Advanced computational methods

d Accessible computing power

Fig. 1 | Key factors driving VLS technology breakthroughs for generation of high-quality hits and leads. **a**, More than 200,000 protein structures in the PDB, plus private collections, have more than 90% of protein families covered with high-resolution X-ray and more recently cryo-electron microscopy structures, often in distinct functional states, with remaining gaps also filled by homology or AlphaFold2 models. **b**, The chemical space available for screening and fast synthesis has grown from about 10^7 on-shelf compounds in 2015 to more than 3×10^{10} on-demand compounds in 2022, and can be rapidly expanded beyond 10^{15} diverse and novel compounds. **c**, Computational methods for VLS include advances in fast flexible docking, modular fragment-based algorithms, DL models and hybrid approaches. **d**, Computational tools are supported by rapid growth of affordable cloud computing, GPU acceleration and specialized chips.

and chemical spaces of drug-like compounds, which can be readily made on-demand, rapidly growing beyond billions of compounds[19], and even larger generative spaces with theoretically predicted synthesizability (Box 1). The third factor involves emerging computational approaches that strive to take full advantage of the abundance of 3D structures and ligand data, supported by the broad availability of cloud and graphics processing unit (GPU) computing resources to support these methods at scale. This includes structure-based virtual screening of ultra-large libraries[20–22], using accelerated[23–25] and modular[26] screening approaches, as well as recent growth of data-driven machine learning (ML) and DL methods for predicting ADMET and PK properties and activities[27].

Although the impacts of the recent structural revolution[17] and computing hardware in drug discovery[28] are comprehensively reviewed elsewhere, here we focus on the ongoing expansion of accessible drug-like chemical spaces as well as current developments in computational methods for ligand discovery and optimization. We detail how emerging computational tools applied in gigaspace can facilitate the cost-effective discovery of hundreds or even thousands of highly diverse, potent, target-selective and drug-like ligands for a desired target, and put them in the context of experimental approaches (Table 1). Although the full impact of new computational technologies is only starting to affect clinical development, we suggest that their synergistic combination with experimental testing and validation in the drug discovery ecosystem can markedly improve its efficiency in producing better therapeutics.

Expansion of accessible chemical space

Why bigger is better

The limited size and diversity of screening libraries have long been a bottleneck for detection of novel potent ligands and for the whole process of drug discovery. An average 'affordable' high-throughput screening (HTS) campaign[29] uses screening libraries of about 50,000–500,000 compounds and is expected to yield only a few true hits after secondary validation. Those hits, if any, are usually rather weak, non-selective, have suboptimal ADMET and PK properties and unknown binding mode, so their discovery entails years of painstaking trial-and-error optimization efforts to produce a lead molecule with satisfying potency and all the other requirements for preclinical development. Scaling of HTS to a few million compounds can be afforded only in big pharma, and it still does not make that much difference in terms of the quality of resulting hits. Likewise, virtual libraries that use in silico screening were traditionally limited to a collection of compounds available in stock from vendors, usually comprising fewer than 10 million unique compounds, therefore the scale advantage over HTS was marginal.

Although chasing the full coverage of the enormous drug-like chemical space (estimated at more than 10^{63} compounds)[30] is a futile endeavour, expanding the screening of on-demand libraries by several orders of magnitude to billions and more of previously unexplored drug-like compounds, either physical or virtual, is expected to change the drug discovery model in several ways. First, it can proportionally increase the number of potential hits in the initial screening[31] (Fig. 2). This abundance of ligands in the library also increases the chances of identification of more potent or selective ligands, as well as ligands with better physicochemical properties. This has been demonstrated in ultra-large virtual screening campaigns for several targets, revealing highly potent ligands with affinities often in the mid-nanomolar to sub-nanomolar range[20–23,26]. Second, the accessibility of hit analogues in the same on-demand spaces streamlines a generation of meaningful structure–activity relationship (SAR)-by-catalogue and further optimization steps, reducing the amount of elaborate custom synthesis. Last, although the library scale is important, properly constructed gigascale libraries can expand chemical diversity (even with a few chemical reactions[32]), chemical novelty and patentability of the hits, as almost all on-demand compounds have never been synthesized before.

Physical libraries

Several approaches have been developed recently to push the library size limits in HTS, including combinatorial chemistry and large-scale pooling of the compounds for parallel assays. For example, affinity-selection mass spectrometry techniques can be applied to identify binders directly in pools of thousands of compounds[33] without the need for labelling. DNA-encoded libraries (DELs) and cost-effective approaches to generate and screen them have also been developed[34], making it possible to work with as many as approximately 10^{10} compounds in a single test tube[35]. These methods have their own limitations; as DELs are created by tagging ligands with unique DNA sequences through a linker, DNA conjugation limits the chemistries possible for the combinatorial assembly of the library. Screening of DELs may also yield a large number of false negatives by blocking important moieties for binding and, more importantly, false positives by nonspecific binding of DNA labels, so expensive off-DNA resynthesis of hit compounds is needed for their validation. To avoid this resynthesis, it has been suggested to use ML modes trained on DEL results for each target to predict drug-like ligands from on-demand chemical spaces, as described in ref. 36.

Virtual on-demand libraries

In silico screening of virtual libraries by fast computational approaches has long been touted as a cost-effective way to overcome the limitations of physical libraries. Only recently, however, have synthetic chemistry

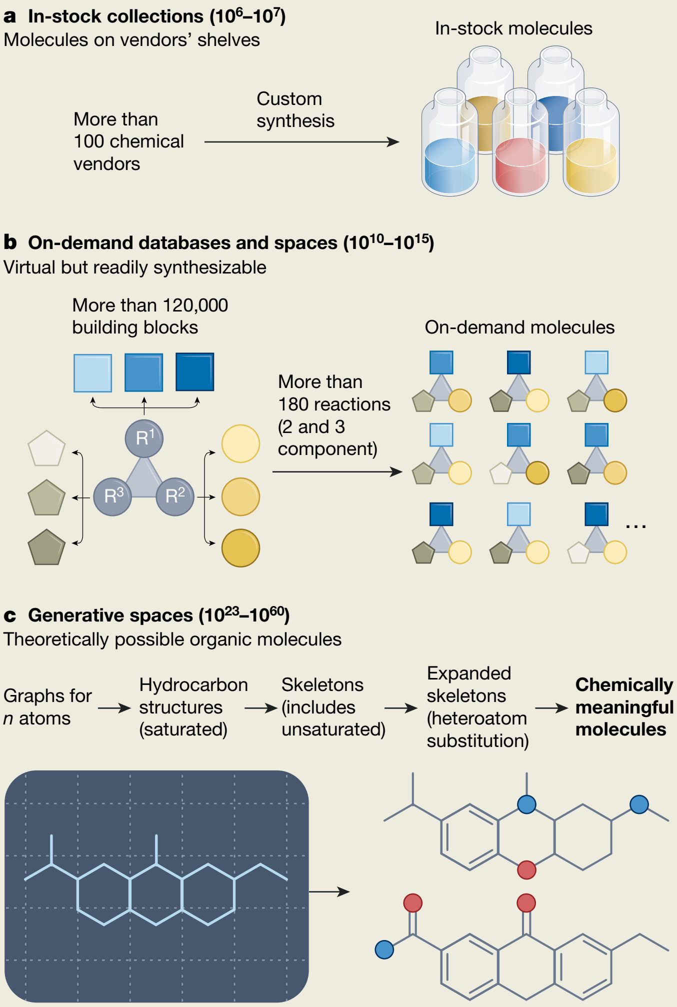
Box 1

Types of chemical libraries and spaces for drug discovery

Pharma companies amass collections of compounds for screening in-house, whereas in-stock collections from vendors (see the figure, part **a**) allow fast (less than 1 week) delivery, contain unique and advanced chemical scaffolds, are easily searchable and are HTS compatible. However, the high cost of handling physical libraries, their slow linear growth, limited size and novelty constrain their applications.

More recently, virtual on-demand chemical databases (fully enumerated) and spaces (not enumerated) allow fast parallel synthesis from available building blocks, using validated or optimized protocols, with synthetic success of more than 80% and delivery in 2–3 weeks (see the figure, part **b**). The virtual chemical spaces assure high chemical novelty and allow fast polynomial growth with the addition of new synthons and reaction scaffolds, including 4+ component reactions. Examples include Enamine REAL, Galaxy by WuXi, CHEMriya by Otava and private databases and spaces at pharmaceutical companies.

Generative spaces, unlike on-demand spaces, comprise theoretically possible molecules and collectively could comprise all chemical space (see the figure, part **c**). Such spaces are limited only by theoretical plausibility, estimated as 10^{23}–10^{60} of drug-like compounds. Although allowing comprehensive space coverage, the reaction path and success rate of generated compounds are unknown, and thus require computational prediction of their practical synthesizability. Examples of generative spaces and their subsets include GDB-13, GDB-17, GDB-18 and GDBChEMBL.



a **In-stock collections (10^6–10^7)**
Molecules on vendors' shelves

b **On-demand databases and spaces (10^{10}–10^{15})**
Virtual but readily synthesizable

c **Generative spaces (10^{23}–10^{60})**
Theoretically possible organic molecules

In-stock molecules
More than 100 chemical vendors
Custom synthesis
More than 120,000 building blocks
On-demand molecules
More than 180 reactions (2 and 3 component)
R1
R2
R3
...
Graphs for n atoms
Hydrocarbon structures (saturated)
Skeletons (includes unsaturated)
Expanded skeletons (heteroatom substitution)
Chemically meaningful molecules

and cheminformatics approaches been developed to break out of these limits and construct virtual on-demand libraries that explore much larger chemical space, as reviewed in refs. 37,38. In 2017, the readily accessible (REAL) database by Enamine[19,39] became the first commercially available on-demand library based on the robust reaction principle[40], whereas the US National Institutes of Health developed synthetically accessible virtual inventory (SAVI)[41], which also uses Enamine building blocks. The REAL database uses carefully selected and optimized parallel synthesis protocols and a curated collection of in-stock building blocks, making it possible to guarantee the fast (less than 4 weeks), reliable (80% success rate) and affordable synthesis of a set of compounds[21]. Driven by new reactions and diverse building blocks, the fully enumerated REAL database has grown from approximately 170 million compounds in 2017 to more than 5.5 billion compounds in 2022 and comprises the bulk of the popular ZINC20 virtual screening database[42]. The practical utility of the REAL database has been recently demonstrated in several major prospective screening campaigns[20,21,23,24], some of them taking further hit optimization steps in the same chemical space, yielding selective nanomolar and even sub-nanomolar ligands without any custom synthesis[20,21]. Similar ultra-large virtual libraries (that is, GalaXi (http://www.wuxiapptec.com) and CHEMriya (http://chemriya.com)) are available commercially, although their synthetic success rates are yet to be published.

Virtual chemical spaces

The modular nature of on-demand virtual libraries supports further growth by the addition of reactions and building blocks. However, building, maintaining and searching fully enumerated chemical libraries comprising more than a few billion compounds become slow and impractical. Such gigascale virtual libraries are therefore usually maintained as non-enumerated chemical spaces, defined by a specific set of building blocks and reactions (or transforms), as comprehensively reviewed in ref. 38. Within pharma, one of the first published examples includes PGVL by Pfizer[37,43], the most recent version of which uses a set of 1,244 reactions and in-house reagents to account for 10^{14} compounds. Other biopharma companies have their own virtual chemical spaces[38,44], although their details are often not in the public domain. Among commercially available chemical spaces, GalaXi Space by WuXi (approximately 8 billion compounds), CHEMriya by Otava (11.8 billion compounds) and Enamine REAL Space (36 billion compounds)[45] are among the largest and most established. In addition to their enormous sizes, these virtual spaces are highly novel and diverse, and have

Table 1 | Comparison of experimentally driven HTS, fragment-based ligand discovery, gigascale DEL screening and gigascale VLS

	HTS	Fragment-based ligand discovery	Gigascale DEL screening	Gigascale VLS
Initial library size	10^5–10^7	10^3–10^5	10^{10}	10^{10}–10^{15}
Hit rate (%)	0.01–0.5	1–5	0.01–0.5	10–40[a]
Expected initial hit affinity	Weak (1–10 μM)	Very weak (100–1,000 μM) small fragments	Medium (0.1–10 μM)	Medium-high (0.01–10 μM)
Further steps to lead identifications	SAR by custom synthesis, QSAR-driven optimization	Merging or growing of fragments, structure-based and QSAR optimization	Label-free hit resynthesis, QSAR-driven optimization with custom synthesis	Extensive SAR-by-catalogue, structure-based and QSAR optimization
Expected number of custom syntheses to lead	500–1,000	500–1,000	200–500	0–50 (mostly on demand or easy parallel synthesis)
Composition of matter patentability	Hits are not novel, need modifications or scaffold hopping to achieve IP novelty	Fragment hits are not novel, require rational design to achieve IP novelty	Depends on the DNA-encoded library	Most hits are not previously synthesized and have IP novelty
Limitations	Modest library size, unknown binding mode, expensive equipment	Expensive NMR, X-ray and BIACORE equipment, many optimization steps	Many false positives, off-DNA resynthesis of hits needed	Computational resources (but reduced more than 1,000 times by modular VLS)

IP, intellectual property. [a]Fraction of predicted candidate hits that were confirmed experimentally.

minimal overlap (less than 10%) between each other[46]. Currently, the largest commercial space, Enamine REAL Space, is an extension to the REAL database that maintains the same synthetic speed, rate and cost guarantees, covering more than 170 reactions and more than 137,000 building blocks (Box 1). Most of these reactions are two-component or three-component, but more four-component or even five-component reactions are being explored, enabling higher-order combinatorics. This space can be easily expanded to 10^{15} compounds based on available reactions and extended building block sets, for example, 680 million of make on demand (MADE) building blocks[47], although synthesis of such compounds involves more steps and is more expensive. To represent and navigate combinatorial chemical spaces without their full enumeration, specialized cheminformatics tools have been developed, from fragment-based chemical similarity searches[48] to more elaborate 3D molecular similarity search methods based on atomic property fields such as rapid isostere discovery engine (RIDE)[38].

An alternative approach proposed to building chemical spaces generates hypothetically synthesizable compounds following simple rules of synthetic feasibility and chemical stability. Thus, the generated databases (GDB) predict compounds that can be made of a specific number of atoms; for example, GDB-17 contained 166.4 billion molecules of up to 17 atoms of C, N, O, S and halogens[49], whereas GDB-18 made up of 18 atoms would reach an estimated 10^{13} compounds[38]. Other generative approaches based on narrower definitions of chemical spaces are now used in de novo ligand design with DL-based generative chemistry (for example, ref. 50), as discussed below.

Although the synthetic success rate for some of the commercial on-demand chemical spaces (for example, Enamine REAL Space) have been thoroughly validated[20–24,26,42], synthetic accessibilities and success rates of other chemical spaces remain unpublished[38]. These are important metrics for the practical sustainability of on-demand synthesis because reduced success rates or unreasonable time and cost would diminish its advantage over custom synthesis.

Computational approaches to drug design

Challenges of gigascale screening

Chemical spaces of gigascale and terrascale, provided that they maintain high drug likeness and diversity, are expected to harbour millions of potential hits and thousands of potential lead series for any target. Moreover, their highly tractable robust synthesis simplifies any downstream medicinal chemistry efforts towards final drug candidates.

Dealing with such virtual libraries, however, calls for new computational approaches that meet special requirements for both speed and accuracy. They have to be fast enough to handle gigascale libraries. If docking of a compound takes 10 s per CPU core, it would take more than 3,000 years to screen 10^{10} compounds on a single CPU core, or cost approximately US $1 million on a computing cloud at the cheapest CPU rates. At the same time, gigascale screening must be extremely accurate, safeguarding against false-positive hits that effectively cheat the scoring function by exploiting its holes and approximations[31]. Even a one-in-a-million rate of false positives in a 10^{10} compound library would comprise 10,000 false hits, which may flood out any hit candidate selection. The artefact rate and nature may depend on the target and screening algorithms and should be carefully addressed in screening and post-processing. Although there is no one simple solution for such artefacts, some practical and reasonably cost-effective remedies include: (1) selection based on the consensus of two different scoring functions, (2) selection of highly diverse hits (many artefacts cluster to similar compounds), (3) hedging the bets from several ranges of scores[31] and (4) manually curating the final list of compounds for any unusual interactions. Ultimately, it is highly desirable to fix as many remaining 'holes in the scoring functions' as possible, and reoptimize them for high selectivity in the range of scores where the top true hits of gigaspace are found. Missing some hits in screening (false negatives) would be well tolerated because of the huge number of potential hits in the 10^{10} space (for example, losing 50% of a million potential hits is perfectly fine), so some trade-off in score sensitivity is acceptable.

The major types of computational approaches to screening a protein target for potential ligands are summarized in Table 2. Below, we discuss some emerging technologies and how they can best fit into the overall DDD pipeline to take full advantage of growing on-demand chemical spaces.

Receptor structure-based screening

In silico screening by docking molecules of the virtual library into a receptor structure and predicting its 'binding score' is a well-established approach to hit and lead discovery and had a key role in recent drug discovery success stories[11,17,51]. The docking procedure itself can use molecular mechanics, often in internal coordinate representation, for rapid conformational sampling of fully flexible ligands[52,53], using empirical 3D shape-matching approaches[54,55], or combining them in a hybrid docking funnel[56,57]. Special attention is devoted to ligand scoring functions, which are designed to reliably remove non-binders to minimize

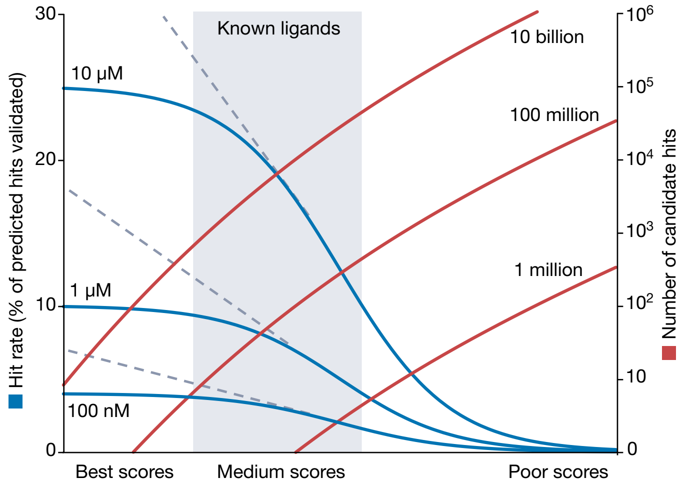


Fig. 2 | Benefits of a bigger chemical space. The red curves in log scale illustrate the distribution of screening hits with binding scores better than *X* for libraries of 10 billion, 100 million and 1 million compounds, as estimated from previous VLS and V-SYNTHES screening campaigns. The blue curves illustrate the approximate dependence of the experimental hit rate on the predicted docking score for 10-µM, 1-µM and 100-nM thresholds[20]. This analysis (semi-quantitative, as it varies from target to target) suggests that screening of more than 100 million compounds lifts the limitations of smaller libraries, extending the tail of the hit distribution towards better binding scores with high hit rates, and allowing for identification of proportionally more experimental hits with higher affinity. Note also two important factors justifying further growth of screening libraries to 10 billion and more: (1) the candidate hits for synthesis and experimental testing are usually picked as a result of target-dependent post-processing of several thousands of top-scoring compounds, which selects for novelty, diversity, drug likeness and often interactions with specific receptor residues. Thus, the more good-scoring compounds that are identified, the better overall selection can be made. (2) Saturation of the hit rate curves at best scores is not a universal rule but a result of the limited accuracy of fast scoring functions used in screening. Using more accurate docking or scoring approaches (flexible docking, quantum mechanical and free energy perturbation) in the post-processing step can extend a meaningful correlation of binding score with affinity further left (grey dashed curves), potentially bringing even more high-affinity hits for gigascale chemical spaces.

false-positive predictions, which is especially relevant with the growth of library size. Blind assessments of the performance of structure-based algorithms have been routinely performed as a D3R Grand Challenge community effort[58,59], showing continuous improvements in ligand pose and binding energy predictions for the best algorithms.

Results of the many successful structure-based prospective screening campaigns have been published over the years covering all major classes of targets, most recently GPCRs, as reviewed in refs. 17,51,60, whereas countless more have been used in industry. The focused candidate ligand sets, predicted by such screening, often show useful (10–40%) hit rates in experimental testing[60], yielding novel hits for many targets with potencies in the 0.1–10-µM range (for those that are published, at least). Further steps in optimization of the initial hits obtained from standard screening libraries of less than 10 million compounds, however, usually require expensive custom synthesis of analogues, which has been afforded only in a few published cases[20,61].

Identification of hits directly in much larger chemical spaces such as REAL Space not only can bring more and better hits[31] but also supports their optimization, as any resulting hit has thousands of analogues and derivatives in the same on-demand space. This advantage was especially helpful for such challenging targets as SARS-CoV-2 main protease (M^{pro}), for which hundreds of standard virtual ligand screening (VLS) attempts came up empty-handed[62] (see discussion on M^{pro} challenges in 'Hybrid in vitro–in silico approaches' below). Although the initial hit rates were low even in the ultra-large screens, VirtualFlow[24] of the REAL database with 1.4 billion compounds still identified hits in the 10–100-µM range, which were optimized via on-demand synthesis[63] to yield quality leads with the best compound Z222979552 (half maximal inhibitory concentration (IC$_{50}$) = 1.0 µM). Another ultra-large screen of 235 million compounds, based on a newer M^{pro} structure with a non-covalent inhibitor (Protein Data Bank (PDB) ID: 6W63), also produced viable hits, fast optimization of which resulted in the discovery of nanomolar M^{pro} inhibitors in just 4 months by a combination of on-demand and simple custom chemistry[64]. The best compound in this work had good in vitro ADMET properties, with an affinity of 38 nM and a cell-based antiviral potency of 77 nM, which are comparable to clinically used PF-07321332 (nirmatrelvir)[65].

With increasing library sizes, the computational time and cost of docking itself become the main bottleneck in screening, even with massively parallel cloud computing[60]. Iterative approaches have been recently suggested to tackle libraries of this size; for example, VirtualFlow used stepwise filtering of the whole library with docking algorithms of increasing accuracy to screen approximately 1.4 billion Enamine REAL compounds[23,24]. Although improving speed several-fold, the method still requires a fully enumerated library and its computational cost grows linearly with the number of compounds, limiting its applicability in rapidly expanding chemical spaces.

Modular synthon-based approaches

The idea of designing molecules from a limited set of fragments to optimally fill the receptor binding pocket has been entertained from the early years of drug discovery, implemented, for example, in the LUDI algorithm[66]. However, custom synthesis of the designed compounds remained the major bottleneck of such approaches. The recently developed virtual synthon hierarchical enumeration screening (V-SYNTHES)[26] technology applies fragment-based design to on-demand chemical spaces, thus avoiding the challenges of custom synthesis (Fig. 3). Starting with the catalogue of REAL Space reactions and building blocks (synthons), V-SYNTHES first prepares a minimal library of representative chemical fragments by fully enumerating synthons at one of the attachment points, capping the other position (or positions) with a methyl or phenyl group. Docking-based screening then allows selection of the top-scoring fragments (for example, the top 0.1%) that are predicted to bind well into the target pocket. This is repeated for a second position (and then third and fourth positions, if available), and the resulting focused libraries are screened at each iteration against the target pocket. At the final step, the top approximately 50,000 full compounds from REAL Space are docked with more elaborate and accurate docking parameters or methods, and the top-ranking candidates are filtered for novelty, diversity and variety of desired drug-like properties. In post-processing, the best 50–500 compounds are selected for synthesis and testing. Our assessment suggests that combining synthons with the scaffolds and capping them with dummy minimal groups in the V-SYNTHES algorithm is a critical requirement for optimal fragment predictions because reactive groups of building blocks and scaffolds often create strong, yet false, interactions that are not present in the full molecule. Another important part of the algorithm is the evaluation of the fragment-binding pose in the target, which prioritizes those hits with minimal caps pointed into a region of the pocket where the fragment has space to grow.

Initially applied to discover new chemotypes for cannabinoid receptor CB$_2$ antagonists, V-SYNTHES has shown a hit rate of 23% for submicromolar ligands, which exceeded the hit rate of standard VLS by fivefold, while taking about 100 times less computational resources[26]. A similar hit rate was found for the ROCK1 kinase screening in the same study, with one hit in the low nanomolar range[26]. V-SYNTHES is being applied to other therapeutically relevant targets with well-defined pocket structures.

Table 2 | Major types of virtual screening algorithms

Type	Approach	Scalability	Applications	Requirements	Examples
Protein structure based	Fast empirical docking	10^6–10^9	Separate ligands from non-binders	High-resolution structures	DOCK[54], GOLD[149], AutoDock[55]
	Molecular mechanics based	10^6–10^8	Separate ligands from non-binders	High-resolution structures	ICM docking[52], ROSETTALigand[53], Glide[56,57]
	Flexible receptor docking	10^3–10^5	Separate ligands from non-binders	Medium-resolution structures	IFD-MD[150]
	Modular VLS	10^{10}–10^{15}	Separate ligands from non-binders	High-resolution structures	V-SYNTHES[26], Chemical Space Docking[151]
	Free energy calculations	10^2–10^3	Affinity ranking	High-resolution structures	FEP+[112], AB-FEP[113,114]
	QM/MM	10^1–10^3	Ion binding, transition state	High-resolution structures	Reviewed in ref. 152
Ligand based	2D/3D QSAR	Up to 10^8	Screening and optimization	Ligand activity large datasets	AutoQSAR[153], APF[154]
	3D pharmacophore and APF screening	Up to 10^{10}	Screening	Ligand activity data	Reviewed in ref. 155, RIDE[98]
	ML/DL-QSAR	Up to 10^{10}	Screening and affinity predictions	Ligand activity large datasets	Q.E.D[78], LSTM-NN[156]
	Chemical space search	Up to 10^{26}	Selection of analogues	Starting ligand (or ligands)	InfiniSee[45]
	QSPR-DL	Up to 10^{10}	Predict solubility, lipophilicity, bioavailability, brain permeability, among others	Large datasets on ligand properties	AstraZeneca PK prediction[73], prediction of oral bioavailability[72–74]
Hybrid	3D interaction fingerprints	Up to 10^{10}	Improved docking and ligand selection	Data on ligand activity and protein–ligand 3D complexes	SIFt[157]
	3D/graph DL	10^6–10^9	Affinity prediction	Data on ligand activity and protein–ligand 3D complexes	Graph-CNN[82,83], 3D-CNN[84,85]
	Dock/DL iterations	10^8–10^{10}	Separate ligands from non-binders	High-resolution structures	MolPal[25], active learning[110], deep docking[111]
	Dock to AI 3D protein models	10^6–10^8	Separate ligands from non-binders	Protein target sequence	AlphaFold[99,100], RosettaFold[101]
	DL-based 3D score function	10^6–10^8	Separate ligands from non-binders	High-resolution structures	RT-CNN[98]

Examples are for illustration only; we apologize for including only a few of the many important programs and tools that are available, due to space limitations. APF, atomic property field; FEP, free energy perturbation; AB-FEP, absolute protein-ligand binding FEP; LSTM-NN, long short-term memory networks-neural networks; SIFt, structural interaction fingerprint; CNN, convolutional neural networks; QM/MM, hybrid quantum mechanics/molecular mechanics; RT-CNN, radial topological CNN; IFD-MD, induced-fit docking molecular dynamics.

A similar approach, chemical space docking, has been implemented by BioSolveIT, so far for two-component reactions[67]. This method is even faster, as it docks individual building block fragments and then enumerates them with scaffolds and other synthons. However, there are trade-offs for the extra speed: docking of smaller fragments without scaffolds is less reliable, and their reactive groups often have dissimilar properties from the reaction product. This may introduce strong receptor interactions that are irrelevant to the final compound and can misguide the fragment selection. This is especially true for cycloaddiction reactions and three-component scaffolds, which need further validation in chemical space docking.

Apart from supporting the abundance, chemical diversity and potential quality of hits, structure-based modular approaches are especially effective in identifying hits with robust chemical novelty, as they (1) do not rely on information for existing ligands and (2) identify ligands that have never been synthesized before. This is an important factor in assuring the patentability of the chemical matter for hit compounds and the lead series arising from gigascale screening. Moreover, thousands of easily synthesizable analogues assure extensive SAR-by-catalogue for the best hits, which, for example, enabled approximately 100-fold potency and selectivity improvement for the CB_2 V-SYNTHES hits[26]. Availability of the multilayer on-demand chemical space extensions (for example, supported by MADE building blocks[47]) can also greatly streamline the next steps in lead optimization through 'virtual MedChem', thus reducing extensive custom synthesis.

Data-driven approaches and DL

In the era of AI-based face recognition, ChatGPT and AlphaFold[68], there is enormous interest in applications of data-driven DL approaches across drug discovery, from target identification to lead optimization to translational medicine (as reviewed in refs. 69–71).

Data-driven approaches have a long history in drug discovery, in which ML algorithms such as support vector machine, random forest and neural networks have been used extensively to predict ligand properties and on-targets activities, albeit with mixed results. Accurate quantitative structure–property relationship (QSPR) models can predict physicochemical (for example, solubility and lipophilicity) and pharmacokinetic (for example, bioavailability and blood–brain barrier penetration) properties, in which large and broad experimental datasets for model training are available and continue to grow[72–74]. ML is also implemented in many quantitative SAR (QSAR) algorithms[75], in which the training set and the resulting models are focused on a given target and a chemical scaffold, helping to guide lead affinity and potency optimization. Methods based on extensive ligand–target binding datasets, chemical similarity clustering and network-based approaches have also been suggested for drug repurposing[76,77].

The advent of DL takes data-driven models to the next level, allowing analysis of much larger and diverse datasets while deriving more complicated non-linear relationships, with vast literature describing specific DL methodologies and applications to drug discovery[27,70]. By its 'learning from examples' nature, AI requires comprehensive ligand

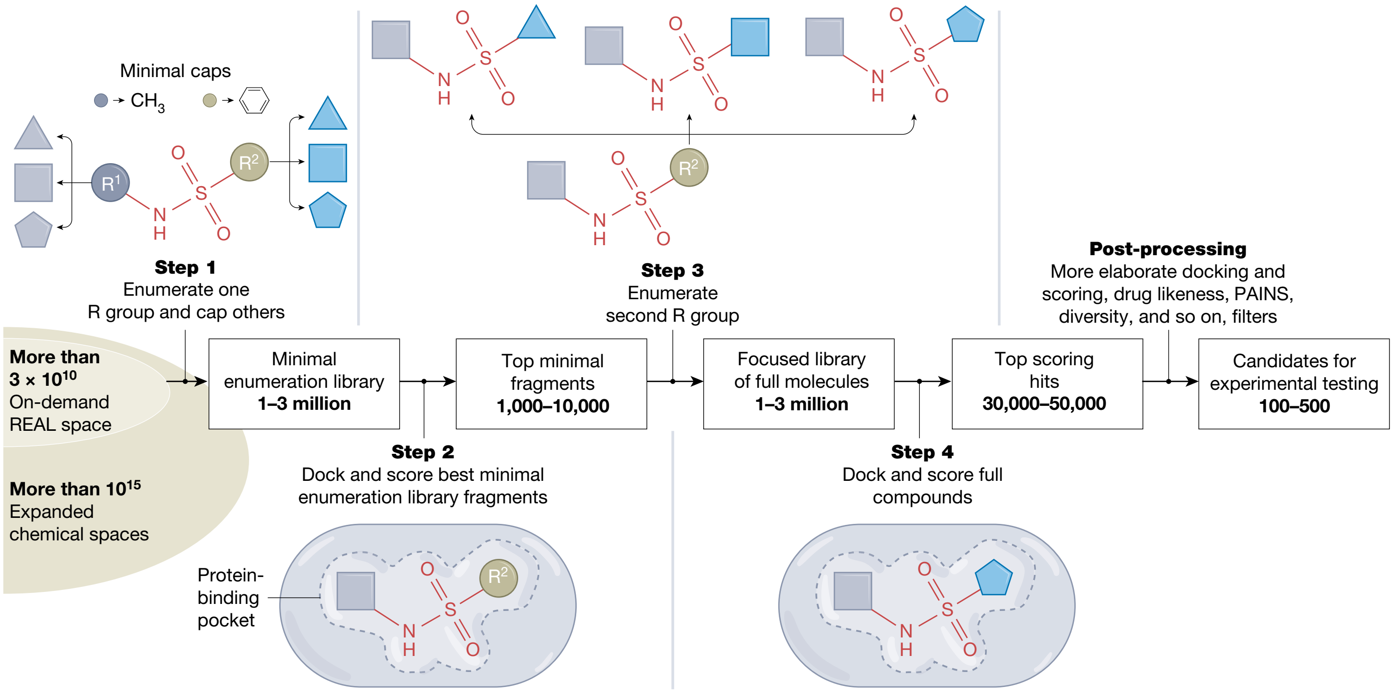

Minimal caps
CH3
R1
R2
N
H
O
S
O
Step 1
Enumerate one R group and cap others
Step 3
Enumerate second R group
Post-processing
More elaborate docking and scoring, drug likeness, PAINS, diversity, and so on, filters
More than 3 × 10^10
On-demand REAL space
More than 10^15
Expanded chemical spaces
Minimal enumeration library
1–3 million
Top minimal fragments
1,000–10,000
Focused library of full molecules
1–3 million
Top scoring hits
30,000–50,000
Candidates for experimental testing
100–500
Step 2
Dock and score best minimal enumeration library fragments
Step 4
Dock and score full compounds
Protein-binding pocket

Fig. 3 | Synthon-based hierarchical screening. An overview of the V-SYNTHES algorithm allowing effective screening of more than 31 billion compounds in REAL Space or even larger chemical spaces, while performing enumeration and docking of only small fractions of molecules. The algorithm, illustrated here using a two-component reaction based on a sulfonamide scaffold with R_1 and R_2 synthons, can be applied to hundreds of optimized two-component, three-component or more-component reactions by iteratively repeating steps 3 and 4 until fully enumerated molecules optimally fitting the target pocket are obtained. PAINS, pan assay interference compounds.

datasets for training the predictive models. For QSPR, large public and private databases have been accumulated, with various properties such as solubility, lipophilicity or in vitro proxies for oral bioavailability and brain permeability experimentally measured for many thousands of diverse compounds, allowing prediction of these properties in a broad range of new compounds.

The quality of QSAR models, however, differs for different target classes depending on data availability, with the most advances achieved for the kinase superfamily and aminergic GPCRs. An unbiased benchmark of the best ML QSAR models was given by a recent IDG-DREAM Drug-Kinase Binding Prediction Challenge with the participation of more than 200 experts[78]. The top predictive models in this blind assessment included kernel learning, gradient boosting and DL-based algorithms. The top-performing model (from team Q.E.D) used a kernel regression, protein sequence similarity and affinity values of more than 60,000 compound–kinase pairs between 13,608 compounds and 527 kinases from ChEMBL[79] and Drug Target Commons[80] databases as the training data. The best DL model used as many as 900,000 experimental ligand-binding data points for training, but still trailed the much simpler kernel model in performance. The best models achieved a Spearman rank coefficient of 0.53 with a root-mean-square error of 0.95 for the predicted versus experimental pK_d values in the challenge set. Such accuracy was found to be on par with the accuracy and recall of single-point experimental assays for kinase inhibition, and may be useful in screenings for the initial hits for less explored kinases and guiding lead optimization. Note, however, that the kinase family is unique as it is the largest class of more than 500 targets, all possessing similar orthosteric binding pockets and sharing high cross-selectivity. The distant second family with systematic cross-reactivity comprises about 50 aminergic GPCRs, whereas other GPCR families and other cross-reactive protein families are much smaller. The performance and generalizability of ML and DL methods for these and other targets remain to be tested.

The development of broadly generalizable or even universal models is the key aspiration of AI-driven drug discovery. One of the directions here is to extract general models of binding affinities (binding score functions) from data on both known ligand activities and corresponding protein–ligand 3D structures, for example, collected in the PDBbind database[81] or obtained from docking. Such models explore various approaches to represent the data and network architectures, including spatial graph-convolutional models[82,83], 3D deep convolutional neural networks[84,85] or their combinations[86]. A recent study, however, found that regardless of neural network architecture, an explicit description of non-covalent intermolecular interactions in the PDBbind complexes does not provide any statistical advantage compared with simpler approximations of only ligand or only receptor that omit the interactions[87]. Therefore, the good performances of DL models based on PDBbind rely on memorizing similar ligands and receptors, rather than on capturing general information about their binding. One possible explanation for this phenomenon is that the PDBbind database does not have an adequate presentation of 'negative space', that is, ligands with suboptimal interaction patterns to enforce the training.

This mishap exemplifies the need for a better understanding of behaviour of DL models and their dependence on the training data, which is widely recognized in the AI community. It has been shown that DL models, especially based on limited datasets lacking negative data, are prone to overtraining and spurious performance, sometimes leading to whole classes of models deemed 'useless'[88] or severely biased by subjective factors defining the training dataset[89]. Statistical tools are being developed to define the applicability range and carefully validate the performance of the models. One of the proposed concepts is the predictability, computability and stability framework for 'veridical data science'[90]. Adequate selection of quality data has been specifically identified by leaders of the AI community as the major requirement for closing the 'production gap', or the inability of ML models to succeed when they are deployed in the real world, thus calling for a data-centric approach to AI[91,92]. There have also been attempts to develop tools to make AI 'explainable', that is, able to formulate some general trends in the data, specifically in the drug discovery applications[93].

Despite these challenges and limitations, AI is already starting to make a substantial effect on drug discovery, with the first AI-based drug candidates making it into the preclinical and clinical studies. For kinases, the AI-driven compounds were reported as potent and effective in vivo inhibitors of the receptor tyrosine kinase DDR1, which is involved in fibrosis[9]. Phase I clinical trials have been announced for ISM001-055 (also known as INS018_055) for the treatment of idiopathic pulmonary fibrosis[10], although the identity of the compound and its target has not been disclosed. For GPCRs, AI-driven compounds targeting 5-HT_{1A}, dual 5-HT_{1A}–5-HT_{2A} and A_{2A} receptors have recently entered clinical trials, providing further support for the AI-driven drug discovery concept. These first success stories are coming from kinase and GPCR families with already well-studied pharmacology, and the compounds show close chemical similarity to known high-affinity scaffolds[94]. It is important for the next generation of DL drug candidates to improve in novelty and applicability range.

Hybrid computational approaches

As discussed above, physics-based and data-driven approaches have distinct advantages and limitations in predicting ligand potency. Structure-based docking predictions are naturally generalizable to any target with 3D structures and can be more accurate, especially in eliminating false positives as the main challenge of screening. Conversely, data-driven methods may work in lieu of structures and can be faster, especially with GPU acceleration, although they struggle to generalize beyond data-rich classes of targets. Therefore, there are numerous ongoing efforts to combine physics-based and data-driven approaches in some synergistic ways in general[95], and in drug discovery specifically[96].

In virtual screening approaches, a synergetic use of physics-based docking with data-based scoring functions may be highly beneficial. Moreover, if the physics-based and data-based scoring functions are relatively independent and both generate enrichment in the selected focused libraries, their combination can reduce the false-positive rates and improve the quality of the hits. This synergy is reflected in the latest 3DR Grand Challenge 4 results for ligand IC_{50} predictions[59], in which the top methods that used a combination of both physics-based and ML scoring outperformed those that did not use ML. Going forward, thorough benchmarking of physics-based, ML and hybrid approaches will be a key focus of a new Critical Assessment of Computational Hit-finding Experiments (CACHE), which will assess five specific scenarios relevant to practical hit and lead discovery and optimization[97].

At a deeper level, the results of accurate physics-based docking (in addition to experimental data, for example, from PDBbind[81]) can be used to train generalized graph or 3D DL models predicting ligand–receptor affinity. This would help to markedly expand the training dataset and balance positive and negative (suboptimal binding) examples, which is important to avoid the overtraining issues described in ref. 87. Such DL-based 3D scoring functions for predicting molecular binding affinity from a docked protein–ligand complex are being developed and benchmarked, most recently RTCNN[98], although their practical utility remains to be demonstrated.

To expand the range of structure-based docking applicability to those targets lacking high-resolution structures, it is also tempting to use AI-derived AlphaFold2 (refs. 99,100) or RosettaFold[101] 3D models, which already show utility in many applications, including protein–protein and protein–peptide docking[102]. Traditional homology models based on close protein similarity, especially when refined with known ligands[103], have been used in small-molecule docking and virtual screening[104], therefore AlphaFold2 is expected to further expand the scope of structural modelling and its accuracy. In a recent report, AlphaFold2 models, augmented by other AI approaches, helped to identify a cyclin-dependent kinase 20 (CDK20) small-molecule inhibitor, although at a modest affinity of 8.9 μM (ref. 105). More general benchmarking of the performance of AlphaFold2 models in virtual screening, however, gives mixed results. In a benchmark focused on targets with existing crystal structures, most AlphaFold2 models had to be cleaned from loops blocking the binding pocket and/or augmented with known ion or other cofactors to achieve reasonable enrichment of hits[106]. For the more practical cases of targets lacking experimental structures, especially for target classes with less obvious structural homologies in the ligand-binding pocket, the performance of AlphaFold2 models in small-molecule docking showed disappointing results in recent assessments for GPCR and antibacterial targets[107,108]. The recently developed AphaFill approach[109] for 'transplanting' small-molecule cofactors and ligands form PDB structures to homologous AlphaFold2 models can potentially help to validate and optimize these models, although further assessment of their utility for docking and virtual screening is ongoing.

To speed up virtual screening of ultra-large chemical libraries, several groups have suggested hybrid iterative approaches, in which results of structure-based docking of a sparse library subset are used to train ML models, which are then used to filter the whole library to further reduce its size. These methods, including MolPal[25], Active Learning[110] and DeepDocking[111], report as much as 14–100 reduction in the computational cost for libraries of 1.4 billion compounds, although it is not clear how they would scale to rapidly growing chemical spaces.

We should emphasize here that scoring functions in fast-docking algorithms and ML models are primarily designed and trained to effectively separate potential target binders from non-binders, although they are not very accurate in predictions of binding affinities or potencies. For more accurate potency predictions, the smaller focused library of candidate binders selected by the initial AI or docking-based screening can be further analysed and ranked using more elaborate physics-based tools, including free energy perturbation methods for relative[112] and absolute[113–115] free energy of ligand binding. Although these methods are much slower, utilization of GPU accelerated calculations[28] holds the potential for their broader application in post-processing in virtual screening campaigns to further enrich the hit rates for high-affinity candidates (Fig. 2), as well as in lead optimization stages.

Future challenges

Further growth of readily accessible chemical spaces

The advent of fast and practical methods for screening gigascale chemical spaces for drug discovery stimulates further growth of these on-demand spaces, supporting better diversity and the overall quality of identified hits and leads. Specifically developed for V-SYNTHES screening, the xREAL extension of Enamine REAL Space now comprises 173 billion compounds[116], and can be further expanded to 10^{15} compounds and beyond by tapping into an even larger building block set (for example, to 680 million of MADE building blocks[47]), by including four-component or five-component scaffolds, and by using new click-like chemistries as they are discovered. Real-world testing of MADE-enhanced REAL Space, and other commercial and proprietary chemical spaces will allow a broader assessment of their synthesizability and overall utility[38,117,118]. In parallel, specialized ultra-large libraries can be built for important scaffolds underrepresented in general purpose on-demand spaces, for example, screening of a virtual library of 75 million easily synthesizable tetrahydropyridines recently yielded potent agonists for the 5-HT_{2A} receptor[119].

Further growth of the on-demand chemical space size and diversity is also supported by recent development of new robust reactions for the click-like assembly of building blocks. As well as 'classical' azide-alkyne cycloaddition click chemistry[120], recognized by the 2022 Nobel Prize in chemistry[121], and optimized click-like reactions including SuFEx[122], more recent developments such as Ni-electrocatalysed doubly decarboxylative cross-coupling[123] show promise. Other carbon–carbon forming reactions use methyliminodiacetic acid boronates for Csp^2–Csp^2 couplings[124], and most recently tetramethyl *N*-methyliminodiacetic

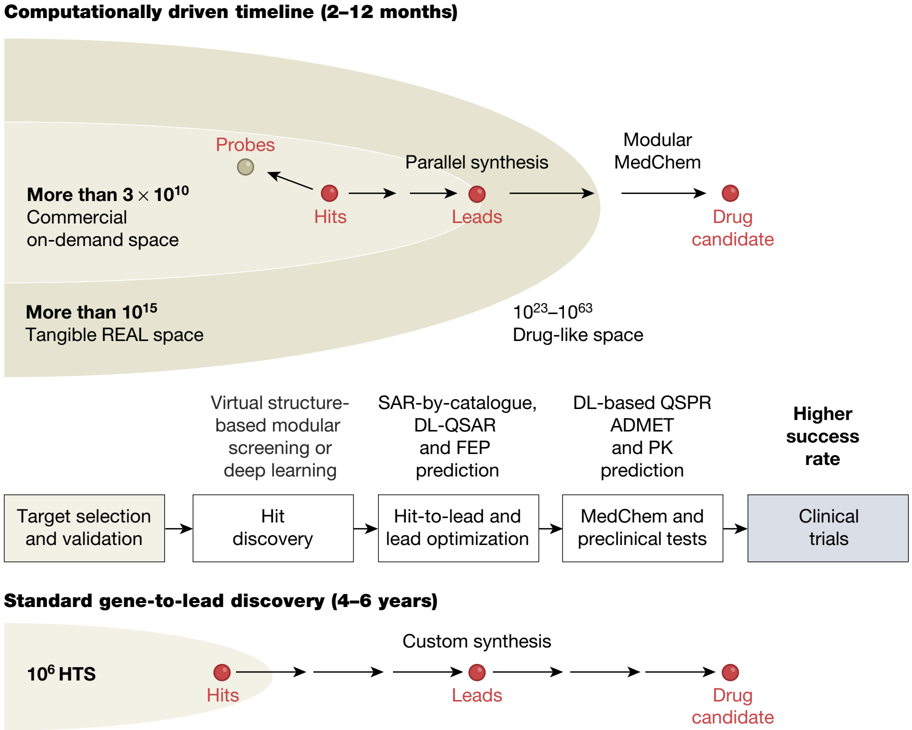

Computationally driven timeline (2–12 months)
Probes
More than 3 × 10^10
Commercial
on-demand space
Hits
Parallel synthesis
Leads
Modular
MedChem
Drug
candidate
More than 10^15
Tangible REAL space
10^23–10^63
Drug-like space
Virtual structure-based modular screening or deep learning
SAR-by-catalogue, DL-QSAR and FEP prediction
DL-based QSPR ADMET and PK prediction
Higher success rate
Target selection and validation
Hit discovery
Hit-to-lead and lead optimization
MedChem and preclinical tests
Clinical trials
Standard gene-to-lead discovery (4–6 years)
10^6 HTS
Hits
Custom synthesis
Leads
Drug
candidate

Fig. 4 | Computationally driven drug discovery. Schematic comparison of the standard HTS plus custom synthesis-driven discovery pipeline versus the computationally driven pipeline. The latter is based on easily accessible on-demand or generative virtual chemical spaces, as well as structure-based and AI-based computational tools that streamline each step of the drug discovery process.

acid boronates[125] for stereospecific C*sp*^{3}–C bond formation. Each of these reactions applied iteratively can generate new on-demand chemical spaces of billions of diverse compounds operating with a limited number of building blocks. Similar to the routinely used automatic assembly of amino acids in peptide synthesis, fully automated processes could be carried out with robots capable of producing a library of drug-like compounds on demand using combinations of a few thousand diverse building blocks[126–128]. Such machines are already working, although scaling-up production of thousands of specialized building blocks remains the bottleneck.

The development of more robust generative chemical spaces can also be supported by new computational approaches in synthetic chemistry, for example, predictions of new iterative reaction sequences[129] or synthetic routes and feasibility from DL-based retrosynthetic analysis[130]. In generative models, synthesizability predictions can be coupled with predictions of potency and other properties towards higher levels of automated chemical design[131]. Thus, generative adversarial networks combined with reinforcement learning (GAN-RL) were recently used to predict synthetic feasibility, novelty and biological activity of compounds, enabling the iterative cycle of in silico optimization, synthesis and testing of the ligands in vitro[50,132]. When applied within a set of well-established reactions and pharmacologically explored classes of targets, these approaches already yield useful hits and leads, leading to clinical candidates[50,132]. However, the wider potential of automated chemical design concepts and robotic synthesis in drug discovery remains to be seen.

Hybrid in vitro–in silico approaches

Although blind benchmarking and recent prospective screening success stories for the growing number of targets support utility of modern computational tools, there are whole classes of challenging targets, in which existing in silico screening approaches are not expected to fare very well by themselves. Some of the hardest cases are targets with cryptic or shallow pockets that have to open or undergo a substantial induced fit to engage ligand, as often found when targeting allosteric sites, for example, in kinases or GPCRs, or protein–protein interactions in signalling pathways.

Although bioinformatics and molecular dynamics approaches can help to detect and analyse allosteric and cryptic pockets[133], computational tools alone are often insufficient to support ligand discovery for such challenging sites. The cryptic and shallow pockets, however, have been rather successfully handled by fragment-based drug discovery approaches, which start with experimental screening for the binding of small fragments. The initial hits are found by very sensitive methods, such as BIACORE, NMR, X-ray[134,135] and potentially cryo-electron microscopy[136], to reliably detect weak binding, usually in the 10–100-μM range. The initial screening of the target can be also performed with fragments decorated by a chemical warhead enabling proximity-driven covalent attachment of a low-affinity ligand[137]. In either case, elaboration of initial fragment hits to full high-affinity ligands is the key bottleneck of fragment-based drug discovery, which requires a major effort involving 'growing' the fragment or linking two or more fragments together. This is usually an iterative process involving custom ligand design and synthesis that can take many years[134,138]. At the same time, structure-based virtual screening can help to computationally elaborate the fragments to match the experimentally identified conformation of the target binding pocket. Most cost-effectively, this approach can be applied when fragment hits are identified from the on-demand space building blocks or their close analogues for easy elaboration in the same on-demand space[139].

The recent examples of hybrid fragment-based computational design approaches targeting SARS-CoV-2 inhibitors highlight the challenges presented by such targets and allow head-to-head comparisons to ultra-large VLS. One of the studies was aimed at the SARS-CoV-2 NSP3 conserved macrodomain enzyme (Mac1), which is a target critical for the pathogenesis and lethality of the virus. Building on crystallographic detection of the low-affinity (180 μM) fragments weakly binding Mac1 (ref. 139), merging of the fragments identified a 1-μM hit, quickly optimized by catalogue synthesis to a 0.4-μM lead[140]. In the same study, an ultra-scale screening of 400 million REAL database identified more than 100 new diverse chemotypes of drug-like ligands, with follow-up SAR-by-catalogue optimization yielding a 1.7-μM lead[140]. For the SARS-CoV-2 main protease M^{pro}, the COVID Moonshot initiative published results of crystallographic screening of 1,500 small fragments with 71 hits bound in different subpockets of the shallow active site, albeit none of them showing in vitro inhibition of protease even at 100 μM (ref. 141). Numerous groups crowdsourcing the follow-up computational design and screening of merged and growing fragments helped to discover several SAR series, including a non-covalent M^{pro} inhibitor with an enzymatic IC_{50} of 21 μM. Further optimization by both structure-based and AI-driven computational approaches, which used more than 10 million MADE Enamine building blocks, led to the discovery of preclinical candidates with cell-based IC_{50} in the approximately 100-nM range, approaching the potency of nirmatrelvir[65]. The enormous scale, urgency and complexity of this Moonshot effort with more than 2,400 compounds synthesized on demand and measured in more than 10,000 assays are unprecedented and this highlights the challenges of de novo design of non-covalent inhibitors of M^{pro}.

Beyond the Moonshot initiative, a flood of virtual screening efforts yielded mostly disappointing results[62], for example, the antimalaria drug ebselen, which was proposed in an early virtual screen[142], failed in clinical trials. Most of these studies, however, screened small-ligand sets focused on repurposing existing drugs, lacked experimental support and used the first structure of M^{pro} solved in a covalent ligand complex (PDB ID: 6LU7) that was suboptimal for docking non-covalent molecules[142].

In comparison, several studies screening ultra-large libraries were able to identify de novo non-covalent M^{pro} inhibitors in the 10–100-μM range[24,62,63,143], while experimentally testing only a few hundred synthesized on-demand compounds. One of these studies further elaborated on these weak VLS hits by testing their Enamine on-demand analogues, revealing a lead with IC_{50} = 1 μM in cell-based assays, and validating its non-covalent binding crystallographically[63]. Another study based on a later, more suitable non-covalent co-crystal structure of M^{pro} (PDB ID: 6W63) used an ultra-large docking and optimization strategy to discover even more potent 38-nM lead compounds[64]. Note that, although the results of the initial ultra-large screenings for M^{pro} were modest, they were on par with the much more elaborate and expensive efforts of the Moonshot hybrid approach, with simple on-demand optimization leading to similar-quality preclinical candidates. These examples suggest that even for challenging shallow pockets, structure-based virtual screening can often provide a viable alternative when performed at gigascale and supported by accurate structures, sufficient testing and optimization effort.

Outlook towards computer-driven drug discovery

With all the challenges and caveats, the emerging capability of in silico tools to effectively tap into the enormous abundance and diversity of drug-like on-demand chemical spaces at the key target-to-hit-to-lead-to-clinic stages make it tempting to call for the transformation of the DDD ecosystem from computer-aided to computer-driven[144] (Fig. 4). At the early hit identification stage, the ultra-scale virtual screening approaches, both structure-based and AI-based, are becoming mainstream in providing fast and cost-effective entry points into drug discovery campaigns. At the hit-to-lead stage, the more elaborate potency prediction tools such as free energy perturbation and AI-based QSAR often guide rational optimization of ligand potency. Beyond the on-target potency and selectivity, various data-driven computational tools are routinely used in multiparameter optimization of the lead series that includes ADMET and PK properties. Of note, chemical spaces of more than 10^{10} diverse compounds are likely to contain millions of initial hits for each target[20] (Box 1), thousands of potent and selective leads and, with some limited medicinal chemistry in the same highly tractable chemical space, drug candidates ready for preclinical studies. To harness this potential, the computational tools need to become more robust and better integrated into the overall discovery pipeline to ensure their impact in translating initial hits into preclinical and clinical development.

One should not forget here that any computational models, however useful or accurate, may never ensure that all of the predictions are correct. In practice, the best virtual screening campaigns result in 10–40% of candidate hits confirmed in experimental validation, whereas the best affinity predictions used in optimization rarely have accuracy better than 1 kcal mol^{-1} root-mean-square error. Similar limitations apply to current computational models predicting ADMET and PK properties. Therefore, computational predictions always need experimental validation in robust in vitro and in vivo assays at each step of the pipeline. At the same time, experimental testing of predictions also provides data that can feed back into improving the quality of the models by expanding their training datasets, especially for the ligand property predictions. Thus, the DL-based QSPR models will greatly benefit from further accumulating data in cell-permeability assays such as CACO-2 and MDCK, as well as new advanced technologies such as organs-on-a-chip or functional organoids to provide better estimates of ADMET and PK properties without cumbersome in vivo experiments. The ability to train ADMET and PK models with in vitro assay data representing the most relevant species for drug development (typically mouse, rat and human) would also help to address species variability as a major challenge for successful translational studies. All of this creates a virtuous cycle for improving computational models to the point at which they can drive compound selection for most DDD end points. When combined with more accurate in vitro testing, this may reduce and eventually eliminate animal test requirements (as recently indicated by FDA)[145].

Building hybrid in silico–in vitro pipelines with easy access to the enormous on-demand chemical space at all stages of the gene-to-lead process can help to generate abundant pools of diverse lead compounds with optimal potency, selectivity and ADMET and PK properties, resulting in less compromise in multiparameter optimization for clinical candidates. Running such data-rich computationally driven pipelines requires overarching data management tools for drug discovery, many of them being implemented in pharma and academic DDD centres[146,147]. Building computationally driven pipelines will also help to reveal weak or missing links, in which new approaches and additional data may be needed to generate improved models, thus helping to fill the remaining computational gaps in the DDD pipeline. Provided this systematic integration continues, computer-driven ligand discovery has a great potential to reduce the entry barriers for generating molecules for numerous lines of inquiry, whether it is in vivo probes for new and understudied targets[148], polypharmacology and pluridimensional signalling, or drug candidates for rare diseases and personalized medicine.

1. Austin, D. & Hayford, T. Research and development in the pharmaceutical industry. *CBO* https://www.cbo.gov/publication/57126 (2021).
2. Sun, D., Gao, W., Hu, H. & Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B* **12**, 3049–3062 (2022).
3. Bajorath, J. Computer-aided drug discovery. *F1000Res.* **4**, F1000 Faculty Rev-1630 (2015).
4. Van Drie, J. H. Computer-aided drug design: the next 20 years. *J. Comput. Aided Mol. Des.* **21**, 591–601 (2007).
5. Talele, T. T., Khedkar, S. A. & Rigby, A. C. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr. Top. Med. Chem.* **10**, 127–141 (2010).
6. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharmacal. Res.* **38**, 1686–1701 (2015).
7. Sabe, V. T. et al. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: a review. *Eur. J. Med. Chem.* **224**, 113705 (2021).
8. Jayatunga, M. K., Xie, W., Ruder, L., Schulze, U. & Meier, C. AI in small-molecule drug discovery: a coming wave. *Nat. Rev. Drug Discov.* **21**, 175–176 (2022).
9. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
This study claims the discovery of a lead candidate in just 21 days, using generative AI, synthesis, and in vitro and in vivo testing of the compounds.

10. US National Library of Medicine. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT05154240#contactlocation (2022).
11. Schrodinger. Schrödinger announces FDA clearance of investigational new drug application for SGR-1505, a MALT1 inhibitor. *Schrodinger* https://ir.schrodinger.com/node/8621/pdf (2022).
    This press release states that combined physics-based and ML methods enabled a computational screen of 8.2 billion compounds and the selection of a clinical candidate after 10 months and only 78 molecules synthesized.
12. Jones, N. Crystallography: atomic secrets. *Nature* **505**, 602–603 (2014).
13. Liu, W. et al. Serial femtosecond crystallography of G protein–coupled receptors. *Science* **342**, 1521–1524 (2013).
14. Nannenga, B. L. & Gonen, T. The cryo-EM method microcrystal electron diffraction (MicroED). *Nat. Methods* **16**, 369–379 (2019).
15. Fernandez-Leiro, R. & Scheres, S. H. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537**, 339–346 (2016).
16. Renaud, J.-P. et al. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).
17. Congreve, M., de Graaf, C., Swain, N. A. & Tate, C. G. Impact of GPCR structures on drug discovery. *Cell* **181**, 81–91 (2020).
18. Santos, R. et al. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2017).
19. Grygorenko, O. O. et al. Generating multibillion chemical space of readily accessible screening compounds. *iScience* **23**, 101681 (2020).
20. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
    This is ultra-large docking study also carefully assessed the advantages and potential pitfalls of expanding chemical space.
21. Stein, R. M. et al. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **579**, 609–614 (2020).
    This study shows ultra-large docking that resulted in subnanomolar hits for a GPCR.
22. Alon, A. et al. Structures of the sigma2 receptor enable docking for bioactive ligand discovery. *Nature* **600**, 759–764 (2021).
23. Gorgulla, C. et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).
    This study shows an iterative library filtering as a first approach to accelerate ultra-large virtual screening.
24. Gorgulla, C. et al. A multi-pronged approach targeting SARS-CoV-2 proteins using ultra-large virtual screening. *iScience* **24**, 102021 (2021).
25. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **12**, 7866–7881 (2021).
    This study introduces acceleration of ultra-large screening by iteratively combining DL and docking.
26. Sadybekov, A. A. et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022).
    This study introduces the modular concept for screening gigascale spaces, V-SYNTHES, and validates its performance on GPCR and kinase targets.
27. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* **119**, 10520–10594 (2019).
28. Pandey, M. et al. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.* **4**, 211–221 (2022).
29. Blay, V., Tolani, B., Ho, S. P. & Arkin, M. R. High-throughput screening: today's biochemical and cell-based approaches. *Drug Discov. Today* **25**, 1807–1821 (2020).
30. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
31. Lyu, J., Irwin, J. J. & Shoichet, B. K. Modeling the expansion of virtual screening libraries. *Nat. Chem. Biol.* https://doi.org/10.1038/s41589-022-01234-w (2023).
32. Tomberg, A. & Boström, J. Can easy chemistry produce complex, diverse, and novel molecules? *Drug Discov. Today* **25**, 2174–2181 (2020).
33. Muchiri, R. N. & van Breemen, R. B. Affinity selection–mass spectrometry for the discovery of pharmacologically active compounds from combinatorial libraries and natural products. *J. Mass Spectrom.* **56**, e4647 (2021).
34. Fitzgerald, P. R. & Paegel, B. M. DNA-encoded chemistry: drug discovery from a few good reactions. *Chem. Rev.* **121**, 7155–7177 (2021).
35. Neri, D. & Lerner, R. A. DNA-encoded chemical libraries: a selection system based on endowing organic compounds with amplifiable information. *Annu. Rev. Biochem.* **87**, 479–502 (2018).
36. McCloskey, K. et al. Machine learning on DNA-encoded libraries: a new paradigm for hit finding. *J. Med. Chem.* **63**, 8857–8866 (2020).
37. Walters, W. P. Virtual chemical libraries. *J. Med. Chem.* **62**, 1116–1124 (2019).
38. Warr, W. A., Nicklaus, M. C., Nicolaou, C. A. & Rarey, M. Exploration of ultralarge compound collections for drug discovery. *J. Chem. Inf. Model.* **62**, 2021–2034 (2022).
    This is a comprehensive review of the history and recent developments of the on-demand and generative chemical spaces.
39. Enamine. REAL Database. *Enamine* https://enamine.net/compound-collections/real-compounds/real-database (2020).
40. Hartenfeller, M. et al. A collection of robust organic synthesis reactions for in silico molecule design. *J. Chem. Inf. Model.* **51**, 3093–3098 (2011).
41. Patel, H. et al. SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci. Data* **7**, 384 (2020).
42. Irwin, J. J. et al. ZINC20-A free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
43. Hu, Q. et al. Pfizer Global Virtual Library (PGVL): a chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb. Sci.* **14**, 579–589 (2012).
44. 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.* **56**, 1253–1266 (2016).
45. Enamine. REAL Space. *Enamine* https://enamine.net/library-synthesis/real-compounds/real-space-navigator (2022).
46. Bellmann, L., Penner, P., Gastreich, M. & Rarey, M. Comparison of combinatorial fragment spaces and its application to ultralarge make-on-demand compound catalogs. *J. Chem. Inf. Model.* **62**, 553–566 (2022).
47. Enamine. Make on-demand building blocks (MADE). *Enamine* https://enamine.net/building-blocks/made-building-blocks (2022).
48. Hoffmann, T. & Gastreich, M. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov. Today* **24**, 1148–1156 (2019).
49. 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.* **52**, 2864–2875 (2012).
50. Vanhaelen, Q., Lin, Y.-C. & Zhavoronkov, A. The advent of generative chemistry. *ACS Med. Chem. Lett.* **11**, 1496–1505 (2020).
51. Ballante, F., Kooistra, A. J., Kampen, S., de Graaf, C. & Carlsson, J. Structure-based virtual screening for ligands of G protein-coupled receptors: what can molecular docking do for you? *Pharmacol. Rev.* **73**, 527–565 (2021).
52. Neves, M. A., Totrov, M. & Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J. Comput. Aided Mol. Des.* **26**, 675–686 (2012).
53. Meiler, J. & Baker, D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* **65**, 538–548 (2006).
54. Lorber, D. M. & Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Protein Sci.* **7**, 938–950 (1998).
55. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
56. Halgren, T. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
57. Friesner, R. A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
58. Gaieb, Z. et al. D3R grand challenge 3: blind prediction of protein-ligand poses and affinity rankings. *J. Comput. Aided Mol. Des.* **33**, 1–18 (2019).
59. Parks, C. D. et al. D3R grand challenge 4: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.* **34**, 99–119 (2020).
60. Bender, B. J. et al. A practical guide to large-scale docking. *Nat. Protoc.* **16**, 4799–4832 (2021).
61. Manglik, A. et al. Structure-based discovery of opioid analgesics with reduced side effects. *Nature* **537**, 185–190 (2016).
62. Cerón-Carrasco, J. P. When virtual screening yields inactive drugs: dealing with false theoretical friends. *ChemMedChem* **17**, e202200278 (2022).
63. Rossetti, G. G. et al. Non-covalent SARS-CoV-2 M^{pro} inhibitors developed from in silico screen hits. *Sci. Rep.* **12**, 2505 (2022).
64. Luttens, A. et al. Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. *J. Am. Chem. Soc.* **144**, 2905–2920 (2022).
    This study compares fragment-based and ultra-large screening-based discovery of lead candidates for the challenging target.
65. Owen, D. R. et al. An oral SARS-CoV-2 M^{pro} inhibitor clinical candidate for the treatment of COVID-19. *Science* **374**, 1586–1593 (2021).
66. Böhm, H.-J. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comput. Aided Mol. Des.* **6**, 61–78 (1992).
67. Beroza, P. et al. Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. *Nat. Commun.* **13**, 6447 (2022).
68. Jumper, J. et al. Applying and improving AlphaFold at CASP14. *Proteins* **89**, 1711–1721 (2021).
69. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
70. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
    This article provides a comprehensive introduction to DL approaches in drug discovery.
71. Elbadawi, M., Gaisford, S. & Basit, A. W. Advanced machine-learning techniques in drug discovery. *Drug Discov. Today* **26**, 769–777 (2021).
72. Bender, A. & Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet. *Drug Discov. Today* **26**, 511–524 (2021).
73. Davies, M. et al. Improving the accuracy of predicted human pharmacokinetics: lessons learned from the AstraZeneca drug pipeline over two decades. *Trends Pharmacol. Sci.* **41**, 390–408 (2020).
74. Schneckener, S. et al. Prediction of oral bioavailability in rats: transferring insights from in vitro correlations to (deep) machine learning models using in silico model outputs and chemical structure parameters. *J. Chem. Inf. Model.* **59**, 4893–4905 (2019).
75. Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
76. Keiser, M. J. et al. Predicting new molecular targets for known drugs. *Nature* **462**, 175–181 (2009).
77. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331 (2016).
78. Cichońska, A. et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nat. Commun.* **12**, 3307 (2021).
79. Gaulton, A. et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).

80. Tang, J. et al. Drug Target Commons: a community effort to build a consensus knowledge base for drug–target interactions. *Cell Chem. Biol.* **25**, 224–229.e222 (2018).
81. Liu, Z. et al. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).
82. Gaudelet, T. et al. Utilizing graph machine learning within drug discovery and development. *Brief. Bioinform.* **22**, bbab159 (2021).
83. Son, J. & Kim, D. Development of a graph convolutional neural network model for efficient prediction of protein–ligand binding affinities. *PLoS ONE* **16**, e0249404 (2021).
84. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Improving detection of protein–ligand binding sites with 3D segmentation. *Sci. Rep.* **10**, 5035 (2020).
85. Jiménez, J., Škalič, M., Martínez-Rosell, G. & De Fabritiis, G. KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.* **58**, 287–296 (2018).
86. Jones, D. et al. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *J. Chem. Inf. Model.* **61**, 1583–1592 (2021).
87. Volkov, M. et al. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *J. Med. Chem.* **65**, 7946–7958 (2022).
88. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
89. Beker, W. et al. Machine learning may sometimes simply capture literature popularity trends: a case study of heterocyclic Suzuki–Miyaura coupling. *J. Am. Chem. Soc.* **144**, 4819–4827 (2022).
90. Yu, B. & Kumbier, K. Veridical data science. *Proc. Natl Acad. Sci. USA* **117**, 3920–3929 (2020).
This perspective article lays a foundation for veridical AI.
91. Ng, A., Laird, D. & He, L. Data-centric AI competition. *DeepLearning AI* https://https-deeplearning-ai.github.io/data-centric-comp/ (2021).
92. Miranda, L. J. Towards data-centric machine learning: a short review. *LJ Miranda* https://ljvmiranda921.github.io/notebook/2021/07/30/data-centric-ml/ (2021).
93. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).
94. Wills, T. AI drug discovery: assessing the first AI-designed drug candidates to go into human clinical trials. *CAS* https://www.cas.org/resources/cas-insights/drug-discovery/ai-designed-drug-candidates (2022).
95. Meng, C., Seo, S., Cao, D., Griesemer, S. & Liu, Y. When physics meets machine learning: a survey of physics-informed machine learning. Preprint at https://doi.org/10.48550/arXiv.2203.16797 (2022).
96. Thomas, M., Bender, A. & de Graaf, C. Integrating structure-based approaches in generative molecular design. *Curr. Opin. Struct. Biol.* **79**, 102559 (2023).
97. Ackloo, S. et al. CACHE (Critical Assessment of Computational Hit-finding Experiments): a public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat. Rev. Chem.* **6**, 287–295 (2022).
This is an important community initiative for comprehensive performance assessment of computational drug discovery methods.
98. MolSoft. Rapid isostere discovery engine (RIDE). *MolSoft* http://molsoft.com/RIDE.html (2022).
99. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
100. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
101. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
102. Akdel, M. A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* **29**, 1056–1067 (2022).
103. Katritch, V., Rueda, M. & Abagyan, R. Ligand-guided receptor optimization. *Methods Mol. Biol.* **857**, 189–205 (2012).
104. Carlsson, J. et al. Ligand discovery from a dopamine D_3 receptor homology model and crystal structure. *Nat. Chem. Biol.* **7**, 769–778 (2011).
105. Ren, F. et al. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel cyclin-dependent kinase 20 (CDK20) small molecule inhibitor. *Chem. Sci.* **14**, 1443–1452 (2023).
106. Zhang, Y. et al. Benchmarking refined and unrefined AlphaFold2 structures for hit discovery. *J. Chem. Inf. Model.* **63**, 1656–1667 (2023).
107. He, X.-h. et al. AlphaFold2 versus experimental structures: evaluation on G protein-coupled receptors. *Acta Pharmacol. Sin.* **44**, 1–7 (2022).
108. Wong, F. et al. Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. *Mol. Syst. Biol.* **18**, e11081 (2022).
109. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).
110. Yang, Y. et al. Efficient exploration of chemical space with docking and deep learning. *J. Chem. Theory Comput.* **17**, 7106–7119 (2021).
111. Gentile, F. et al. Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat. Protoc.* **17**, 672–697 (2022).
112. Schindler, C. E. M. et al. Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.* **60**, 5457–5474 (2020).
113. Chen, W., Cui, D., Abel, R., Friesner, R. A. & Wang, L. Accurate calculation of absolute protein–ligand binding free energies. Preprint at https://doi.org/10.26434/chemrxiv-2022-2t0dq-v2 (2022).
114. Khalak, Y. et al. Alchemical absolute protein–ligand binding free energies for drug design. *Chem. Sci.* **12**, 13958–13971 (2021).
115. Cournia, Z. et al. Rigorous free energy simulations in virtual screening. *J. Chem. Inf. Model.* **60**, 4153–4169 (2020).
116. xREAL Chemical Space, *Chemspace*, https://chem-space.com/services#v-synthes (2023).
117. Rarey, M., Nicklaus, M. C. & Warr, W. Special issue on reaction informatics and chemical space. *J. Chem. Inf. Model.* **62**, 2009–2010 (2022).
118. Zabolotna, Y. et al. A close-up look at the chemical space of commercially available building blocks for medicinal chemistry. *J. Chem. Inf. Model.* **62**, 2171–2185 (2022).
119. Kaplan, A. L. et al. Bespoke library docking for 5-HT_{2A} receptor agonists with antidepressant activity. *Nature* **610**, 582–591 (2022).
120. Krasiński, A., Fokin, V. V. & Sharpless, K. B. Direct synthesis of 1,5-disubstituted-4-magnesio-1,2,3-triazoles, revisited. *Org. Lett.* **6**, 1237–1240 (2004).
121. The Nobel Prize in Chemistry. *nobelprize.org*, https://www.nobelprize.org/prizes/chemistry/2022/summary/ (2022)
122. Dong, J., Sharpless, K. B., Kwisnek, L., Oakdale, J. S. & Fokin, V. V. SuFEx-based synthesis of polysulfates. *Angew. Chem. Int. Ed. Engl.* **53**, 9466–9470 (2014).
123. Zhang, B. et al. Ni-electrocatalytic C*sp*^{3}-C*sp*^{3} doubly decarboxylative coupling. *Nature* **606**, 313–318 (2022).
124. Gillis, E. P. & Burke, M. D. Iterative cross-couplng with MIDA boronates: towards a general platform for small molecule synthesis. *Aldrichimica Acta* **42**, 17–27 (2009).
125. Blair, D. J. et al. Automated iterative C*sp*^{3}–C bond formation. *Nature* **604**, 92–97 (2022).
This study provides a chemical approach for automation of the C–C bond formation in small-molecule synthesis.
126. Li, J. et al. Synthesis of many different types of organic small molecules using one automated process. *Science* **347**, 1221–1226 (2015).
127. Trobe, M. & Burke, M. D. The molecular industrial revolution: automated synthesis of small molecules. *Angew. Chem. Int. Ed.* **57**, 4192–4214 (2018).
128. Bubliauskas, A. et al. Digitizing chemical synthesis in 3D printed reactionware. *Angew. Chem. Int. Ed.* **61**, e202116108 (2022).
129. Molga, K. et al. A computer algorithm to discover iterative sequences of organic reactions. *Nat. Synth.* **1**, 49–58 (2022).
130. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
131. Goldman, B., Kearnes, S., Kramer, T., Riley, P. & Walters, W. P. Defining levels of automated chemical design. *J. Med. Chem.* **65**, 7073–7087 (2022).
132. Grisoni, F. et al. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Sci. Adv.* **7**, eabg3338 (2021).
133. Wagner, J. R. et al. Emerging computational methods for the rational discovery of allosteric drugs. *Chem. Rev.* **116**, 6370–6390 (2016).
134. Davis, B. J. & Hubbard, R. E. in *Structural Biology in Drug Discovery* (ed. Renaud, J.-P.) 79–98 (2020).
135. de Souza Neto, L. R. et al. In silico strategies to support fragment-to-lead optimization in drug discovery. *Front. Chem.* **8**, 93 (2020).
136. Saur, M. et al. Fragment-based drug discovery using cryo-EM. *Drug Discov. Today* **25**, 485–490 (2020).
137. Kuljanin, M. et al. Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. *Nat. Biotechnol.* **39**, 630–641 (2021).
138. Muegge, I., Martin, Y. C., Hajduk, P. J. & Fesik, S. W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* **42**, 2498–2503 (1999).
139. Schuller, M. et al. Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. *Sci. Adv.* **7**, eabf8711 (2021).
140. Gahbauer, S. et al. Iterative computational design and crystallographic screening identifies potent inhibitors targeting the Nsp3 macrodomain of SARS-CoV-2. *Proc. Natl Acad. Sci. USA* **120**, e2212931120 (2023).
This article demonstrates the application of both hybrid fragment screening-and-merging design and ultra-large library screening to a challenging viral target.
141. Achdout, H. et al. Open science discovery of oral non-covalent SARS-CoV-2 main protease inhibitor therapeutics. Preprint at https://doi.org/10.1101/2020.10.29.339317 (2022).
142. Jin, Z. et al. Structure of M^{pro} from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020).
143. Ton, A. T., Gentile, F., Hsing, M., Ban, F. & Cherkasov, A. Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. *Mol. Inform.* **39**, e2000028 (2020).
144. Frye, L., Bhat, S., Akinsanya, K. & Abel, R. From computer-aided drug discovery to computer-driven drug discovery. *Drug Discov. Today Technol.* **39**, 111–117 (2021).
145. Wadman, M. FDA no longer needs to require animal tests before human drug trials. *Science*, https://doi.org/10.1126/science.adg6264 (2023).
146. Stiefl, N. et al. FOCUS—development of a global communication and modeling platform for applied and computational medicinal chemists. *J. Chem. Inf. Model.* **55**, 896–908 (2015).
147. Schrodinger. LiveDesign. *Schrodinger* https://www.schrodinger.com/sites/default/files/general_ld_rgb_080119_forweb.pdf. (accessed 5 April 2023)
148. Müller, S. et al. Target 2035—update on the quest for a probe for every protein. *RSC Med. Chem.* **13**, 13–21 (2022).
149. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein–ligand docking using GOLD. *Proteins* **52**, 609–623 (2003).
150. Miller, E. B. et al. Reliable and accurate solution to the induced fit docking problem for protein–ligand binding. *J. Chem. Theory Comput.* **17**, 2630–2639 (2021).
151. Chemical space docking. *BioSolveIT* https://www.biosolveit.de/application-academy/chemical-space-docking/ (2022).
152. Cavasotto, C. N. in *Quantum Mechanics in Drug Discovery* (ed. Heifetz, A.) 257–268 (Springer, 2020).
153. Dixon, S. L. et al. AutoQSAR: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling. *Future Med. Chem.* **8**, 1825–1839 (2016).
154. Totrov, M. Atomic property fields: generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. *Chem. Biol. Drug Des.* **71**, 15–27 (2008).
155. Schaller, D. et al. Next generation 3D pharmacophore modeling. *WIREs Comput. Mol. Sci.* **10**, e1468 (2020).

156. Chakravarti, S. K. & Alla, S. R. M. Descriptor free QSAR modeling using deep learning with long short-term memory neural networks. *Front. Artif. Intell.* **2**, 17 (2019).

157. Deng, Z., Chuaqui, C. & Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **47**, 337–344 (2004).

Acknowledgements We thank A. Brooun, A. A. Sadybekov, S. Majumdar, M. M. Babu, Y. Moroz and V. Cherezov for helpful discussions.

Author contributions All authors contributed to the writing of the manuscript.

Competing interests The University of Southern California are in the process of applying for a patent application (no. 63159888) covering the V-SYNTHES method that lists V.K. as a co-inventor.

Additional information

Correspondence and requests for materials should be addressed to Vsevolod Katritch.

Peer review information *Nature* thanks Alexander Tropsha and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.