https://doi.org/10.1038/s44386-025-00031-4

Navigating structure-based drug discovery with emerging innovations in physics- and knowledge-based approaches

Check for updates

Jordy Homing Lam[1,2] ✉ & Vsevolod Katritch[1,2,3,4] ✉

Structure-based drug design is rapidly evolving, driven by advances in both physics-based and knowledge-based methods. These computational approaches are increasingly integrated across all stages of drug discovery. Despite remarkable progress, challenges remain in achieving accuracy, generalizability, computational efficiency, and chemical synthesizability. In this review, we provide a critical overview of advances, strengths, and limitations of recent methods. We also discuss synergies between the two concepts that hold promises for future advancements towards their practical applicability.

Computational approaches have become increasingly involved in all the steps of drug discovery, from target identification to chemical library design and hit discovery and optimization[1–7]. Structure-based virtual screening (SBVS) grounded in physical principles has long been an essential part of drug discovery[8–10]. Recently, major breakthroughs have been achieved in scale and practical utility[11–13]. In parallel, the accumulation of ligand data and the advances in machine learning approaches have improved predictions of compound properties, including synthesizability[3], physicochemical[14,15] and pharmacokinetics[16], and in some cases, target binding and activity, showing utilities in lead optimizations[2], drug repurposing[17], and discovery of new hits. Pure ligand-based approaches, however, require extensive target/domain-specific bioactivity data, available only for targets with already thoroughly studied pharmacology[18–21]. Because drugs act by specifically binding to their biological targets, a detailed structural understanding of these interactions will therefore benefit ligand-target activity prediction and generalization across chemotypes and target classes[22,23]. Computational structure-based drug discovery, in general, uses atomistic high-resolution structural information on the target, whether derived experimentally from crystallographic and cryo-EM data[24], or from modeled structures[25,26], to predict ligand binding poses and affinities (Table 1), and prioritize chemical compounds for synthesis and experimental testing. At the initial hit discovery stage, computational methods are employed to screen for a set of likely binders from vast diversity libraries or chemical spaces too large to be handled experimentally. The hits found at this stage are expected to have activities in the sub-micromolar to high-micromolar range. At the hit-to-lead optimization stage, goals then shift to the development of quantitative structure-activity relationship (QSAR) models, balancing potency and pharmacological properties, with compounds refined by structure and data-guided medicinal chemistry approaches[27].

Computational methods in structure-based drug discovery can be broadly divided into physics-based and knowledge-based approaches. Physics-based methods aim to accurately estimate the binding free energy (BFE) via simulations with electrostatic, Van der Waals, and other physical terms parameterized in closed forms using higher-level ab initio (e.g., quantum mechanics) data. However, in practice, particularly for molecular docking approaches, empirical terms like hydrogen bonding, pi-pi, and hydrophobic interactions are often introduced to approximate true quantum and entropic effects[28]. The scoring functions are also composed of physical forms fitted against benchmark data, often referring to experimental binding affinity data. Active research in physics-based methods continues, with improved force fields[29–33], scoring functions[28,34], and applications of free energy perturbation theories[35–37]. On the other hand, the rapidly growing collection of structural and bioactivity datasets are opening opportunities to develop sophisticated knowledge-based models capable of handling diverse targets and ligand chemotypes[24,38,39]. Knowledge-based models may be interpreted as statistical potentials[40], with parameterized putative energy terms approximating the log likelihood of the inputs, solely derived from data made of empirical or experimental sources (e.g. observed binding geometry, binding affinities, et cetera). The output of a knowledge-based scoring function then represents the significance of a binding interaction on the observed distribution. However, as demonstrated by several recent assessments, knowledge-based methods are doped with specific

[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, University of Southern California, Los Angeles, CA, USA. [3]Department of Chemistry, University of Southern California, Los Angeles, CA, USA. [4]Department of Pharmacology, University of Southern California, Los Angeles, CA, USA. ✉e-mail: homingla@usc.edu; katritch@usc.edu

Table 1 | Glossary of common terms

Glossary	Explanatory Remarks
Apo	A protein in its unbound form, lacking a ligand in the pocket targeted by drug discovery.
Holo	A protein with a ligand in the pocket targeted by drug discovery.
Pocket	A specific region on the target protein, usually a cavity, where a ligands can specifically and selectively bind through non-covalent or covalent interactions and modulate biological activity.
Ligand	A molecule that binds specifically to a protein target through non-covalent or covalent interactions. Ligands can be endogenous or synthetic compounds (e.g., drugs or probes) designed to interact with a target to modulate its function or activity.
Hit discovery	The initial phase in the drug discovery to identity active compounds in high-throughput screening, fragment-based screening or virtual screening.
Hit-to-lead optimization	Chemical modification of the identified hits to improve their binding and functional activity, selectivity, pharmacokinetic profile, and reduce off-target effects.
Retrospective Benchmark	An assessment of methods using historical data with known outcomes, useful in ongoing optimization and validation of prediction methods.
Blind Benchmark	An assessment where the ground truth is known only to the organizers but withheld from participants.
Prospective Benchmark	An assessment, where experimental results are generated after computational prediction, imitating the real drug discovery process.
3D pharmacophore	A spatial arrangement of most important steric and electronic features representing a molecule or a series of molecule.

challenges[41–43]. First, these potentials are prone to deviations in ligand geometry, stereochemistry and steric clashes, which physics-based potentials explicitly guard against. Secondly, unlike linearly regressed physics-based potentials and scoring functions that have a very limited number of parameters fitted, these knowledge-based methods can invoke neural networks with millions of parameters. The overfitting issue, unfortunately, was masked by contaminations among data splits in common benchmarking datasets[41,44–48], which often leads to inflated apparent generalizability of the methods. Therefore, efforts in software and algorithmic development should be balanced with a better understanding of the underlying physical phenomena and of the strengths and weaknesses of datasets.

In all regards, while computational structure-based drug discovery has gained mainstream interest in both pharma and academia, it is undergoing a major change of paradigm. We argue that the merits of computational methods can be dissected into four criteria of success, namely Accuracy, Scalability, Synthesizability, and Generalizability (Fig. 1). Ideally, the best methods should predict highly potent, synthesizable drug-like compounds within reasonable time and memory requirements, considering the costs of computing; they should also be able to generalize across unencountered scenarios absent in the training data. This review will focus on some recent progress in physics-based and knowledge-based approaches, striving towards these goals. In the following sections, we will first discuss the scope of data available, then move on to contextualize aspects affecting the accuracy, scalability, and synthesizability of methods. Finally, we will also touch on several emerging solutions to combine physics- and knowledge-based methodologies in a synergistic manner.

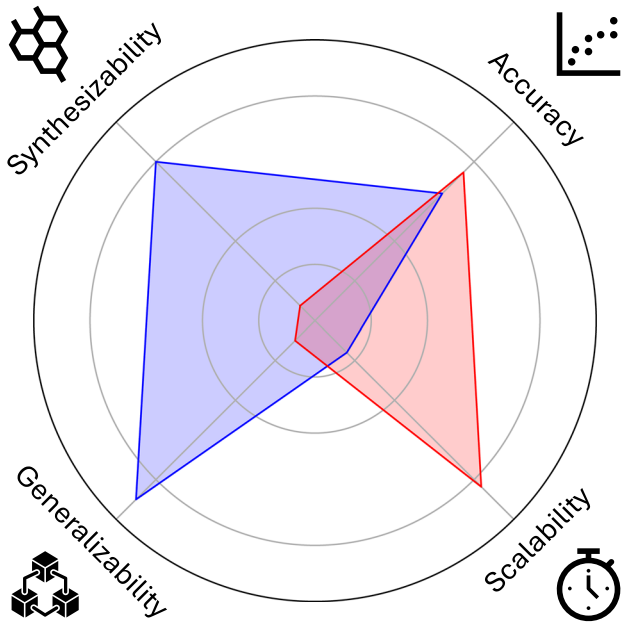

Fig. 1 | The criteria for success in the development of structure-based drug discovery software, namely Accuracy, Scalability, Synthesizability and Generalizability, are depicted as four axes. Two hypothetical computational methods (red and blue triangles) are depicted. The method in red produces potent virtual chemicals with high scalability, but the predicted virtual chemicals are not synthesizable. The method in blue is more resource-demanding and thus less scalable, but its outputs are relatively accurate and highly synthesizable. While the method in red is slightly higher in accuracy within its domain of training data, it suffers in generalizability when being applied to an unencountered scenario.

Data and Benchmarks

Sources of data are primarily structural and bioactivity data, either compiled and integrated from literature and patents, or generated on scale through systematic efforts[49] (Table 2). In physics-based methods, these datasets are mostly reserved for benchmarking and validation, with exceptions in molecular docking approaches, while in knowledge-based methods, they are the sole foundational resources for model training. Benchmarking of methods on these data can be retrospective, blind, or prospective. In this section, we will introduce some of these data sources and discuss steps taken by the community to expand their diversity and to reduce their intrinsic biases. We will also discuss the evolving formats of retrospective and prospective benchmarks in recent years.

Structural data

The largest volume of publicly available structural data is curated by the Protein Data Bank (PDB)[24], though many companies also maintain their own structural biology programs, keeping most results private. The PDB atomic coordinates are interpretations from experimental data, mostly produced by crystallography or cryogenic electron microscopy (cryo-EM). Quality metrics, such as resolution, are explicitly declared in PDB entries. However, other relevant information, such as protonation[50], charges[51], alternative sidechain conformations[52,53], and unassigned densities usually will require additional interpretation[54]. For instance, cryo-EM maps determined at near-atomic (1.9–2.5 Å) global resolution may contain parts resolved with much lower local quality due to the intrinsic flexibility or radiation damage[55–57]. Practically, the reliability of physics-based methods is highly contingent on the accuracy of binding pocket modeling. As a result, significant effort is typically devoted to assessing the quality of modeled pockets at the outset of any drug discovery campaign[58]. Several indicators, such as the temperature factor and more recently LIVQ score[55,57] and DAQ

Table 2 | Examples of databases relevant to structure-based drug discovery

Database (Year Updated)	Data Content	Estimated Number of Relevant Data Entries	Reference
PDB (2025)	Experimental structural data with systematic validation	~30 thousand ligand-target complexes	24
ChEMBL (2025)	Experimental activity data from literature and donated datasets	~24 million records for ~2.8 million compounds	38,39,63,64
BindingDB (2024)	Experimental activity data from literatures and patents	~2.9 million records for ~1.3 million compounds, where ~1.4 million records are distinguished from ChEMBL	62
PDBBind (2020)	PDB structural data and ChEMBL activity data	~24 thousand, ~5500, ~300 ligand-target complexes in the 'General Set', 'Refined Set' and 'Core Set' respectively.	70,73,94
CrossDocked (2020)	Structural data provided by cross-docking	~22.5 million poses, positive and negative examples, generated from ~14 thousands ligand and ~3000 pockets. The cross-validation folds were produced on clusters of target sharing >40% sequence similarity.	45
PoseBusters (2024)	Selected PDB structural data	308 unique protein–ligand complexes from PDB, each containing a unique ligand, stratified by various levels of sequence similarity	41
PLINDER (2024)	PDB structural data with extensive annotations	~449 thousand protein-ligand interaction entries, including cofactors, lipids, and sugars, stratified by various levels of sequence similarity	42

score[59] may reflect on these local qualities. These indicators play a critical role in deciding whether a given protein structure is suitable for downstream modeling and can influence both the initiation of a structure-based project and the inclusion or exclusion of particular datapoints from training sets or virtual screening campaigns. Besides, individual structures in PDB provide only one snapshot of an otherwise quite dynamic molecular system; a collection of structures for the same or similar proteins in different complexes and functional states may also give insights into ligand binding site flexibility and induced fit effects[60]. (See below section "Molecular induced-fit and co-folding").

Bioactivity data

Ligand binding or bioactivity data are compiled from literature, patents, and donated datasets in several public databases, including PubChem[61], ChEMBL[39] and BindingDB[62]. As of its latest 2025 release, ChEMBL contains over 24.2 million bioactivity records for 2.8 million unique compounds, making it the largest publicly available repository of bioactivity[38,63]. It includes quantitative bioactivity data for ligands and their corresponding targets, such as K_i, K_d, IC_{50}, or EC_{50} values, manually extracted from literature, and presented in the 'pChEMBL' column in a standardized negative logarithmic scale. Additional compounds can be found in patent databases, such as SureChEMBL[64] and BindingDB[62], although patents often omit activity data for ligands and conflate active and inactive compounds. When working with such datasets, caution should be applied in integrating activity data from different assays, laboratories, conditions, modes of action, et cetera, which are subjected to variation and noise, e.g., for ChEMBL, it is advised to use only entries with high Confidence Scores, focusing on K_i data, or implement additional custom filters[65,66]. Despite these caveats, these databases provide invaluable sources of ligand activity data for different proteins. Within companies, the public datasets are further complemented by proprietary in-house data. Notably, federated learning approaches have recently enabled training on these confidential datasets without necessitating data disclosure among houses[67]. At the same time, recognition of scarcity and limited quality of ligand binding data has led recently to several commercial and public initiatives to generate ligand binding data for diverse targets systematically and at scale, e.g., Target 2035[68].

Overlaps between structural and bioactivity data

Currently, about 24,000 structural data in the PDB are associated with activity data for ligands in ChEMBL[69]. This count naturally excludes certain entities, such as metabolites, ions, solvents, and water, that ubiquitously exist in PDB but are not considered specific or drug-like ligands. Structural and bioactivity data are integrated in specialized databases, such as PDBbind[70], where the overlapping data are systematically categorized by quality and diversity filters. For instance, in PDBbind 2020, a "Refine Set" (~5500 complexes) was selected from the 'General Set' (~ 23,000 complexes) by applying quality filters on structural and bioactivity records; the "Core Set" complexes (~300 complexes) are representatives from sequence similarity clusters[71,72] of the "Refined Set". The "Core Set" has been employed as the primary test set in the CASF benchmark for evaluating various scoring functions[48]. Additionally, patent records on BindingDB are also juxtaposed with ~10,000 structures available in the PDB. Notably, many structures are not resolved with protein sequences exactly matching the bioactivity records due to mutations or orthologs. In ~5000 such cases, BindingDB provides structures with sequence identity greater than 85% to the proteins used in assays, although this may introduce some additional errors, e.g., due to mutations affecting ligand activity.

Cross-validation practices

Ligand activities are correlated among targets sharing similar structures. As a result, a realistic benchmark, intended to assess generalization on truly unencountered targets, should have similar targets grouped and completely segregated between training and test sets. This data stratification practice and the ambiguous measure of similarity may be derived from metadata, such as sequence similarity clusters[71,72] curated on the PDB. As expected, and broadly discussed by the community, the absence of proper data stratification between training and test sets can adversely overstate performance and generalization capacity in machine learning. For instance, it was shown in Kramer et al.[44] that while a method trained on the "Refined Set" of PDBbind achieves a Pearson correlation of 0.77 when tested on the "Core Set", its performance can drop to 0.46 when proteins with 90% BLAST sequence similarities were excluded from the training set. The nested similarity between "Core Set" and "Refined Set" is obvious, as the "Core Set" was selected from representatives of the "Refined Set", therefore, simple and commonly used index-splitting practices (e.g., excluding exactly overlapping entries from the "Core Set") allow major information leakage from training to test sets. Indeed, recognizing this issue, the PDBbind authors have conducted a thorough investigation and provided a standardized training/test split considering pocket similarity[73]. Recently, some of these aspects were re-examined in the context of cross-fold validations. For instance, in CrossDocked 2020, clustered cross-validation folds referencing Kramer et al.[44] were adopted by grouping targets with over 40% sequence identity into distinct clusters, which were then randomly assigned to folds for 3-fold cross-validation. Similar practices can be found in earlier research[44,47,74]. This assures that targets with sequence similarity (i.e., a high fraction of identical residues after multiple sequence alignment) are not shared across training/testing sets, assuming that pocket similarity is partially covered by sequence identity. However, as ligand-target interaction can occur at the interface of segment(s) of biopolymer(s), data stratification by constructing, pruning and clustering similarity graphs of binding site/

sequence(s) can be a very involved, but worthwhile, exercise[44,45,60,73,75]. Recently, the PLINDER-PL50 dataset[42] has been announced to contain extensively annotated PDB structural data comprising approximately 449,000 protein-ligand interaction systems, including common cofactors and sugars (e.g., ATP and NAG). These additional data can be useful for model training, but caution should be exercised, as ligands that are not drug-like are abundant in the dataset. The PLINDER database also provides correspondence among holo, apo and predicted structures, facilitating analysis of induced-fit inference scenarios. The PLINDER-PL50 dataset[42] is constructed by stratifying on pocket similarity clusters, accounting sequence similarity for residues within the holo pocket after multiple sequence alignment with MMSeqs2[72] or Foldseek[76], whichever is higher. Most recently, in Runs N' Poses[77], a subset of PLINDER, stratified by a combined ligand-pocket similarity metric, was introduced as a test set, showing that a strong dependence of prediction success on similarity to known ligand-receptor pairs persists even for state-of-the-art co-folding methods like AlphaFold3 and Boltz. These developments are positive steps towards a debiased assessment of methodologies, with the recognition that further major community efforts are needed to achieve robust benchmarking of binding pose and affinity predictions[43].

Negative data

Most applications are tasked, first and foremost, with separating binders from non-binders. However, negative data, i.e., data on compounds that do not bind to or exhibit activity against a target, often go undocumented and thus remain scarce in records. Apparently, non-binders are not resolved in high-resolution experimental structures of the receptor pocket either. This absence of negative data creates a significant label imbalance, posing challenges for the training and evaluation of structure-based computational models. For benchmarking applications aimed at virtual ligand screening, especially screening ultra-large and giga-scale libraries[11,13,78], negative examples must be curated. A common strategy for augmenting negative data, pioneered by the Directory of Useful Decoys (DUD) family of datasets[79–81] involves selecting decoys for each target randomly from a large chemical pool of compounds that have physicochemical properties similar but topologically dissimilar to the positive examples. For instance, DUD-E contains 22,886 active compounds targeting 102 proteins from ChEMBL, each paired with 50 property-matched decoys drawn from ZINC[82]. The criteria for decoy selection were designed to counteract biases in scoring functions for physicochemical properties, such as molecular weight, solubility, net charge, number of rotatable bonds, and hydrogen bond donors/acceptors. Recently, DUD-Z addressed additional biases, such as those related to scaffold diversity[80]. These datasets were widely used for benchmarking target models in computational programs[83–85], and more recently, several knowledge-based approaches[86,87]. DUD also explores the cross-docking hypothesis[88], which postulates on the scarce specificity of ligands to only a few targets, even when binding pockets share similarities. This idea was later pursued in the CrossDocked2020 dataset[45] where supposedly active ligands sourced from PDBbind were cross-docked against their non-cognate targets using Smina[89]. The dataset comprises ~22.5 million poses, almost evenly split between cross-docked examples and defective binding poses. This presents efforts to relieve the scarcity of negative structural data with a computational surrogate. Notably, the cross-docking hypothesis can also be harnessed in reverse screening or target fishing scenarios[48,90,91], which evaluates a scoring function on their ability to retrieve known binding targets from a background of largely unrelated proteins, using the ligand as the query. This helps to guard against confirmation bias, or colloquially the "yes-man" problem, where a model consistently favors expected outcomes without genuine discriminative power. Such reverse screening can be useful in drug repurposing and in explaining ligand selectivity and polypharmacology. A more detailed comparison of methods for reverse screening power was discussed in the reference[48].

Retrospective, blind, and prospective benchmarks

Efforts in developing better benchmarks, including retrospective, blind, and prospective assessments, are essential in efforts to further improve computational practices. Retrospective benchmarks discussed above, e.g., PDBBind, BindingDB, and PLINDER, are based on previously published experimental data, where all information is released to developers to enable fast ongoing evaluation/optimization of model performance. Blind benchmarks are organized assessments where ground truths are known to organizers but withheld from participants. Examples include Drug Design Data Resource Grand Challenge (D3R)[92], Comparative Assessment of Scoring Functions (CASF)[48,93,94] GPCR Dock[95] (GPCRD). Finally, prospective benchmarks are assessments on targets and ligands for which no experimental results exist at the time of prediction. An example of truly prospective assessments is Critical Assessment of Computational Hit-finding Experiments (CACHE)[4]. Performances of computational methods are usually assessed in terms of their (1) "docking power" (ability to match with experimental binding pose), (2) "scoring power" (ability to correlate linearly with experimental binding affinity), (3) "ranking power" (ability to reproduce relative ranking of ligands according to their experimental binding affinities) and (4) "screening power" (ability to discriminate true binders from random molecules), with standard metrics and statistical tests[48,93,94,96,97]. Challenges in mastering these powers are well-pronounced in community-wide assessments. For instance, in the latest D3R Grand Challenge 4 concluded in 2020, participants were asked to make predictions on 154 ligands targeting beta-secretase 1 (BACE1)[92]. BACE1 is well known for its vast array of structural (> 300 co-crystals at the time of assessment) and activity data; the given ligands were also known—but undisclosed—to vary in activities ranging from nM to μM. Despite so, results are mixed, with the best methods achieving docking power within 0.5 angstrom in root-mean-square deviation (RMSD) but much less impressive ranking and scoring power with Kendall's τ less than 0.4 and root-mean-square error (RMSE) exceeding 1.3 kcal/mol. The recently introduced CACHE assessments[4,98] play a pivotal role in benchmarking computational methods, as they closely mimic the pipeline of modern drug discovery for given targets. CACHE aims at assessing methods for the prospective identification of new hits and follow-up compounds from vast commercially accessible chemical libraries, such as Enamine REAL and ZINC20, containing billions of compounds. The predicted compounds, 100 from each participating group, are synthesized and tested in vitro, and the best validated hits can be optimized in a round of hit expansion. So far, results of the first four CACHE assessments have been published targeting the WD40 repeat domain of LRRK2 (#1), RNA binding site (#2) and macrodomain (#3) of the NSP13 helicase of SARS-CoV-2, and epigenetic target SETDB1 (#4), with experimental part of CACHE #5-6 ongoing, and CACHE #7 announced. The hit rates in the first four assessments were low, even for the most successful methods, and the best compounds were found in micromolar or even > 10 μM range, reflecting challenging targets selected for CACHE to explore frontier problems. In particular, hit rates for the best methods in CACHE #1 and #2 are still rather low at 3.7% and 9% respectively. Interestingly, results across D3R, CASF, and CACHE all indicated that hybrid workflows, incorporating knowledge-based and physics-based approaches, consistently achieve the most successful outcomes[5,48,86], as will be discussed later.

Accuracy

Physics-based approaches are grounded in the concept of BFE (BFE, ΔG), reflecting the "strength" of binding as a statistical average in some idealized conditions[99]. The value of ΔG can be related to the experimental measurement of binding affinity (dissociation constant K_d) in that $\Delta G = RT \ln(K_d)$ equates in a two-state system (bound/unbound) at a defined temperature T, where $\Delta G = \Delta H - T\Delta S$ is made of enthalpy H and entropy S. Notably, experimental affinities that differ by several thousand folds (e.g., nM versus μM) would imply substantial differences in the BFE, with each ten-fold change corresponding to 1.36 kcal/mol in ΔG at room temperature. Fundamentally, the distributive nature of these terms would require some forms of energy evaluation and stochastic sampling, where convergence is

hard to achieve, so it is often approached by some rough approximations. For instance, molecular docking scoring functions, which comprise physics-based interaction terms, such as electrostatics, van der Waals, bond deformation, and torsional strain, often have their weights fitted against experimental binding affinities. However, it is well known that the sum decomposition of free energies into interaction terms is at best a first-order approximation[100,101]. As a result, while these scores may be useful for ranking compounds, any accuracy claims that equate docking outputs with rigorous free energy computations should be treated with caution. On the other hand, knowledge-based methods or scoring functions can be defined completely independently of physical context or functional forms, solely relying on distributions learned from the data. For instance, they may use structure-affinity datasets or, more commonly, binding poses alone to estimate the fitness of a given pose in specified pocket coordinates. In recent years, knowledge-based methods have also enabled the sampling of binding poses based on a trained knowledge-based function. This may help to soften limitations in the closed-form physics-based scoring function for effects that are unknown or difficult to model. This section will focus on factors affecting the accuracy of physics-based scoring methods and some recent developments in knowledge-based methods.

Evaluation of energy

The gold standards for energy evaluation are quantum mechanical (QM) methods utilizing coupled cluster (CC) theory, arguably followed by Møller–Plesset perturbation theory (MP) and density functional theory (DFT)[102,103]. While the high cost of these methods has limited their applicability to systems with fewer than 100 atoms, c.f. a typical pocket with 400–600 atoms, QM methods remain indispensable "ground truths" for development of force fields. (n.b., highly simplified QM methods are not necessarily more accurate than force fields[104].) The motivation for developing these closed-form functions akin to classical interpretation is the reduction of computational cost from staggering $O(N^7)$ to $O(N^2)$ or even O(N) as in methods for electrostatics calculations[105,106]. Most major force fields, for example, MMFF94[30], OPLS3[29], GAFF[107] and CHARMM[108] were fitted against MP data, OPLS3e[109] was fitted against DFT data. Many of these molecular mechanics (MM) force fields were also tested against CC level data. For instance, OPLS3 was compared on the Genentech torsion benchmark[110], achieving RMSE of 1.02 kcal/mol[111]. Beyond distinction in qualities of the QM data, there are certain aspects in force field construction that are subjected to artisanship[112,113], concerning atom-typing, torsion-typing, partial charge modeling, virtual site assignment, and treatment of polarizability[31,109,112–114]. Some quantum effects are essential in modeling noncovalent interactions observed in medicinal chemistry, such as sigma-holes, edge-to-face and aromatic stacking[115–117]. Besides, torsional barriers can exhibit correlations across neighboring groups[118] and are influenced by long-range electronic effects[119–121], in particular by attractive dispersion in Van der Waals interactions, making it challenging to capture accurate torsion profiles using a simple Fourier expansion of a predefined torsion type. To address these limitations, empirical corrections were introduced in grid-based spline fitting methods, such as the CMAP correction in CHARMM[118], or ICMFF approaches[31,122,123] to refine torsional potentials. However, these methods would generally require case-by-case, conformation-by-conformation QM data, and in some cases experimental data, such as NMR coupling, for refitting each specific system. These approaches, though, are primarily applicable to highly repetitive and debatably well-behaved polymers, such as peptides, where consistent torsional behavior can be systematically parameterized.

Recently, the rapid rise in computing resources and growing availability of ab initio data have led to the development of neural network potentials (NNPs), such as ANI[124] and MACE[125], which aim to enhance the approximation of energies and forces with more explicit physical functional forms parametrized using non-linearly activated neural networks rather than linear weights. Beyond the difference in parametrization methods and possibly applications of variational fitting approaches or other machine learning techniques[126–128], the NNPs also differ from traditional force fields in that the in-situ geometry, chemical elements, and net charges are used as the functional input to derive parameters in an end-to-end differentiable manner to replace the pre-determined atom-typing/torsion-typing exercise. This differs from the traditional force field in that the parameters in traditional force fields are invariant to geometry after training or post-hoc decisions. Different NNPs also vary in the representation of atomic environments to describe in situ geometry[129]. For instance, ANI[124] and AIMNET[127,130] leverage the Behler-Parrinello atom-centered symmetry function[131] to encode local structural information, characterizing interatomic distances and three-body angles within a neighborhood of 4.6A and 3.1 A, respectively. In MACE[15,125], tensor products and contractions in real spherical harmonics are directly used to describe body-order features up to a user-defined degree to assimilate many-body features, serving similar roles to the use of torsion angles in traditional force fields. A key advantage of the atomic cluster expansion[132] (ACE) framework used in MACE is that its inference cost scales linearly with the number of neighboring atoms, in contrast to the combinatorial scaling of an unrefined multipole expansion. Moreover, ACE is systematically improvable in the limit of body-order and polynomial basis expansion, offering a highly expressive representation of atomic interactions[132]. Both of these methods have reached milestones of sub-kcal/mol accuracy, reportedly surpassing force fields, such as GAFF and OPLS[29,107,109], and have been successfully applied in tasks, such as strain energy calculations and molecular dynamics simulations. Notably, concepts in multiscale computation[133], such as QM/MM[134], which apply high-accuracy QM methods to chemically critical regions (e.g., the ligand) while treating surrounding environments with lower-cost MM methods, have been extended in several variants, with additional modeling techniques, such as delta-learning, that replace the MM contribution (as in QM/MM-NN[135] and HDNNP[136]) and/or the QM contribution (as in ANI/EFP[119] and ML/MM[137]) with NN potentials. These advancements aim to generalize energy calculations for chemically diverse or previously uncharacterized systems while maintaining computational efficiency. However, the sensitivity of QM/MM methods to hyperparameters, such as charge models at the QM/MM boundary and numerical quadrature settings in DFT integration, must be carefully assessed to enhance reliability and standardization[138].

Evaluation of binding free energy

The estimation of the BFE ΔG can be performed by applying the free energy perturbation (FEP) formula of Zwanzig[139,140] or separately by applying the thermodynamic integration (TI) formula of Kirkwood[141] and possibly many other methods, such as MM/PBSA[142–144]. In these approaches, the fundamental idea is to exploit the fact that the free energy is a state function, allowing one to compute the BFE between two end states via some intermediate states, without numerically confronting the partition functions[145,146]. As a result, most FEP and TI exercises require design of thermodynamic cycles (i.e. design of intermediate states with a goal to empirically facilitate convergence) and running a series of molecular dynamics (MD) or monte-carlo (MC) simulations to estimate differences in potential energies[145]. The two end states are typically the bound state (ligand staying in the pocket) and the unbound state (ligand in the bulk solvent). The intermediate states can be non-physical, for instance, implemented by gradually morphing (perturbing) force-field parameters and interaction energies to convert between the two end states. These intermediates form a network, which also refers to as the FEP map. Practical FEP calculations can be broadly categorized into computation of absolute BFE (ABFE) and relative BFE (RBFE). While ABFE compares the raw BFE (i.e., ΔG) between bound and unbound states for each ligand on the same target (and may potentially extend to comparison on different targets), the RBFE focuses on the relative differences in BFE (i.e., ΔΔG) among pairs of ligands. In ABFE, the thermodynamic cycle is designed to cover multiple gradual changes in interaction parameters for a single ligand in bulk solvent and pocket, with additional steps to account for spatial

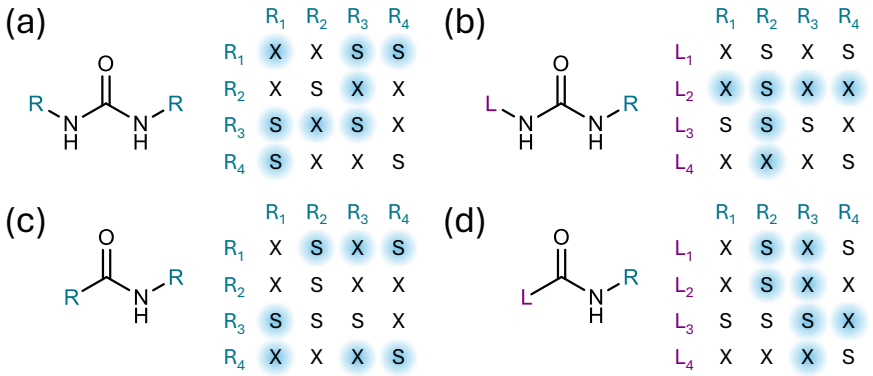

Fig. 2 | Four hypothetical scenarios with synthetic traps. Synthetic traps are indicated by "X". Synthetically accessible products are indicated by "S". Shades indicate products predicted to be synthesizable with general reaction rules or synthesizability restraints. **a** Symmetric attachment points with symmetric vocabulary of functional groups. **b** Symmetric attachment points with asymmetric vocabulary of functional groups. **c** Asymmetric attachment points with symmetric vocabulary of functional groups. **d** Asymmetric attachment points with symmetric vocabulary of functional groups.

restraints (see Fig. 2 from reference[147]). This can involve a huge perturbing step when the electrostatic and Van der Waals interactions involving the whole ligand in bulk solvent and in the pocket are turned off. In contrast, the RBFE is not computed by subtracting individual ABFEs, but by redesigning a thermodynamic cycle that involves two ligands (see Fig. 1 from reference[148]) such that the perturbation steps only concern switching between the mutating part of the ligands instead of swapping entire ligands to enhance convergence. Hence, in RBFE, the FEP map is commonly chosen within a chemically similar congeneric series for the purpose of hit optimization. Current FEP usage is often restricted to FEP maps of several hundred congeneric compounds due to the costly simulations with explicit solvents and membrane. Additionally, for ABFE, the sampling effort is also substantiated by the differences between the end states, assuming no major induced fit from ligand binding[149]. Therefore, the majority of FEP applications are focused on RBFE usage. Besides, it is also worth noting that ABFE and RBFE require accurate knowledge of the end states and the assumption that the binding location and pose, commonly obtained from experiments or molecular docking, are largely preserved among ligands[6,150]. Previously, Schindler et al. reported retrospective and prospective RBFE data using the FEP+ pipeline[151] on 19 in-house targets, where 13 of them show acceptable scoring power (RMSE < 1.5 kcal/mol or Kendall's $\tau > 0.5$)[6]. They also identified subtleties in force-field specifications as a major determinant in accuracy and anticipated that improvement in these dependencies will further benefit accuracy. More recently, Ross et al.[35] re-examined the maximally attainable accuracy in RBFE predictions by comparing FEP+ performance against the reproducibility of experimental assays across a broad chemical and biological landscape. Unlike earlier efforts, which excluded rare or narrow-ranged congeneric series (e.g. macrocycles, charge-altering modifications, and buried water displacement events) to minimize statistical noise, they deliberately retained such data to maximize chemical diversity and alchemical transformation coverage. This inclusive strategy allowed them to evaluate FEP+ on previously underrepresented benchmarks. To evaluate the best accuracy of FEP+ achievable in a perfect setup, a large variety of protomeric, tautomeric, rotameric, and ambiguous binding states were explicitly included in the FEP map. This study concluded that the FEP+ can deliver predictive accuracy approaching the empirical limits of experimental reproducibility (discussed in refs. 65,66), typically around 1 kcal/mol, when best practices in structure preparation and model setup are rigorously followed. To conclude, FEP offers a powerful, theoretical framework for estimating binding free energies in structure-based drug discovery. With carefully constructed thermodynamic cycles and optimized simulation protocols, especially within RBFE framework, these methods can approach the predictive accuracy of experimental assays. However, limitations on force-field accuracy, sampling convergence, and the reliability of binding pose assumptions still pose major bottlenecks. Addressing these challenges will further extend the utility and reliability of FEP-based methods. However, high computational cost limits FEP applications to tens or hundreds of molecules and relies on known binding poses of ligands.

Molecular docking with physics-based approaches

In practice, computational hit discovery tasks have been predominantly driven by molecular docking approaches, which are based on sampling ligand conformations to optimize scoring functions approximating ligand-target interactions. Classical scoring functions are physics-based[8–10], meaning that the scoring function subsumes physical functional forms e.g., electrostatic and Van der Waals energies, although they can be reweighed empirically and refitted to experimental data to achieve specific goals. In general, a molecular docking protocol begins with interpretation of the target and ligand structures, followed by minimization or Monte-Carlo sampling of ligand conformations on a trained scoring function. The outcomes are binding poses associated with scores. While these scores maybe trained on experimental binding affinity data, they are, in general, very rough proxies to the actual ΔG, at best a first-order approximation[100], with many empirical corrections, e.g., hydrogen-bonding terms[28], to soften imperfections in target structure or sampling procedures. In many algorithms, such as DOCK[152], GOLD[153], and Glide[154], target-ligand conformational sampling is limited to translational and rotational degrees of freedom (DOFs) of the ligand, using a set of low-energy ligand conformers precalculated separately. Other methods implemented e.g., in Molsoft ICM[155], AutoDock[156] and RosettaLigand[157] perform explicit sampling of torsional variables in flexible ligands, formulating the problem as a global optimization challenge. In some cases, the entropic contributions due to flexibility in target/ligand conformation and solvent interactions are accounted only implicitly by solvent screening and desolvation terms, e.g., through the generalized Born/surface area model (GB/SA)[158,159]. The efficiency in sampling is maintained through parallelization and implementations of metaheuristic optimization techniques[160], including simulated annealing[156], biased probability Monte Carlo[161], iterative local search[28,154,155] and genetic algorithms[153]. In programs, such as VirtualFlow[12], a series of docking algorithms can be applied with a stepwise increase of computational intensity at different stages of the VS to progressively filter for more promising compounds. Many of these technologies are successfully deployed in retrospective benchmarks and prospective screening campaigns. However, as illustrated by results from the CACHE initiative, prospective hit rates remain modest, highlighting the ongoing challenges in bridging docking predictions with experimental success.

Rescoring with knowledge-based approaches

Recently, knowledge-based innovations were introduced to existing and new molecular docking protocols[86,162–164]. For instance, in GNINA[86,164], candidate flexible ligand poses are first identified through AutoDock Vina with an iterative Monte Carlo search and are then re-ranked by knowledge-

based scores deduced through an ensemble of convolutional neural networks (CNNs). The inputs to the CNNs are atomic coordinates of the complex encoded as Gaussian-like densities on a 3D grid. These CNNs are trained to classify binding poses based on their deviation from co-crystal structures (e.g., above or below a 2 Å threshold), and to predict binding affinities using data from PDBbind. GNINA is ranked among the top-performing methods in the CACHE challenge #1[165]. One limitation of GNINA's voxel-based CNN approach is that rotating the binding pocket can lead to different discretization ('gridding') and hence different predictions. To address this, the training of GNINA is augmented with random rotations[164], which mitigates but does not fully resolve the issue. This limitation can be overcome by including only rotationally invariant features (e.g., radial distribution of atoms, distances and angles)[166] or by carefully incorporating tensorial interactions and produce rotationally equivariant outputs (e.g., through implementations, such as equivariant graph neural network[167]). In the latter approach, it also allows prediction of vectors in addition to scalars.

Molecular docking with knowledge-based approaches

The sampling of ligand conformation, as presented in the molecular docking paradigm, may be accomplished by a multitude of simulation methods running on a trained scoring function, where the objective is to find several local minima. Recent developments in diffusion models have led to a plethora of new molecular docking methods based on knowledge-based scoring functions. The working principle of a diffusion model is to train a neural network to learn the gradient of the distribution through a denoising mechanism. In inference, this learned gradient is used to guide particle sampling dynamics, replacing the costly explicit gradient evaluations from physical potential as done in other metaheuristic alternatives. In DiffDock[162], a diffusion model trained to infer actions (rotation, translation, and torsion), through the use of an equivariant graph neural network[167], to reproduce ligand poses starting from a random pose and, separately, a 'confidence model' is trained to discriminate poses with RMSD criteria, similar to GNINA. The guided sampling is conditioned on a coarse-grained protein structure and chemical identities. This differs from regression-based predecessors, such as EquiBind[168], which did not employ sampling and produced a singular binding pose. However, as shown in the PoseBusters study[41], knowledge-based potentials can often produce unphysical binding poses. While applying a physics-based force-field minimization step may partially correct such artifacts, inconsistencies often persist, reflecting a mismatch between learned statistical potential and physical energy landscapes. This underscores the challenge of using knowledge-based potentials to approximate the rich multimodal physical space of binding poses. In later works, such as SurfDock[169], the input conditions are further enriched with pretrained embeddings, such as sequence embeddings (via ESM-2) and physicochemical features (e.g., electrostatics, hydropathy, shape index, curvature) on the target surface enabled by MaSIF[170]. This multimodal fusion enables more informed sampling over ligand placement within pockets. In a prospective screening campaign targeting ALDH1B1, SurfDock successfully identified seven novel scaffolds with sub-micromolar to low-micromolar affinities, demonstrating its promise as a knowledge-based docking platform with real-world translational capacity.

Molecular docking on modeled target structures

With the advent of predictive modeling methods, such as AlphaFold2 (AF2)[25,171], RoseTTAFold[172] and ESMFold[173], apo structures for most proteins are now readily available. This naturally raises the question of whether any of the SBVS methods, physics- or knowledge-based, can be applied to these computational models. Mixed success has been observed so far. In Karelina et al.[174], several physics-based methods, such as Glide SP, Glide XP, and Rosetta docking, were applied to AF2 models concerning 18 GPCRs (17 class A, one class B). They found that while apo AF2 models captured the binding pocket structure more accurately than apo homology models, the accuracy of the docked ligand poses was not significantly higher than that of the homology model, and is much lower than when they are docked to experimentally determined holo structures of the same ligand. At the same time, Lyu et al.[175] found that, for two high-confidence AF2 models (σ_2 receptor and 5-HT2A receptor), prospective docking using DOCK was as effective as using experimental structures, leading to similar hit rate and ligand quality, although the compounds discovered with AF2 models had mixed functional activities. These results suggest possibilities for using modeled target structures in SDBB, although their practical performance for lower-confidence targets lacking experimental structural characterization remains to be explored.

Molecular induced-fit and co-folding

Ultimately, structure-based modeling strives to accurately predict optimal conformations of target–ligand complexes, taking into account full conformational flexibility of the target. Most drug discovery campaigns use holo structures derived from complexes with endogenous ligands or apo structures that feature well-defined and relatively static pockets, typically limited to minimal sidechain rearrangements and "breathing motions"[176–179]. However, in many other pharmacologically relevant cases, the pocket conformation is unknown, induced by the ligand or formed at protein–protein interfaces. These cryptic, dynamic, and often transient pockets pose special challenges for structure-based drug discovery and are often considered undruggable in general. Simulating such pockets remains difficult due to their intrinsically long timescales and rare-event nature. Enhanced sampling strategies like metadynamics and relaxation-based MD protocols[179–184] can be applied in an attempt to discover hidden pockets, however, their potential for predicting ligand binding to such targets is limited.

Recent advances in knowledge-based structure prediction opened a possibility of directly modeling protein–ligand complexes from sequence and ligand information, allowing the pocket conformation to deviate from known structural templates. For instance, AlphaFold 3 presents a unified diffusion-based architecture that models complexes involving ligands and biopolymers, eliminating reliance on predefined torsion parametrizations or explicit stereochemical corrections[26]. Building on this direction, models, such as NeuralPLexer and Umol adopt generative or dual-mode prediction frameworks to infer atomic complexes from protein sequences and ligand graphs, with or without pocket annotations[185,186]. In a recent preprint, Boltz-2[187] extends this diffusion paradigm further by introducing thermodynamic steering. This technique imposes empirical restraints, such as distance and local geometry priors, by injecting flat-bottom potentials and several soft energy terms into the diffusion process, thereby biasing the sampled structures toward physically realistic configurations. While useful for correcting clashes and restoring valid stereochemistry, the core idea of incorporating energy-like terms into the sampling distribution is conceptually well-established in docking methods. In all, the distinguishing feature of these methods lies in their learned co-folding capability and the integration of structure and affinity prediction. Recent assessments, however, show that current co-folding algorithms largely rely on memorizing poses found for similar ligands in similar targets, fail when ligand-target pairs lack such similarity[77], and produce physically unrealistic predictions in adversarial examples where the pocket residues are all mutated to glycines or phenylalanines[188]. These shortcomings remain to be addressed in future versions of co-folding algorithms, and may require much more experimental data to achieve broad generalizability[68].

Scalability

Computational tools should be balanced with the practical needs of hit discovery. In a typical SBVS campaign, millions and billions of chemical compounds are sorted through to select tens to hundreds of potential hit candidates. Until recently, the screening libraries were limited to a few million in-stock compound libraries[189], from vendors like ChemBridge, ChemDiv, Enamine, Molport, or public databases like ZINC[82,190–192]. This has recently changed with the introduction of on-demand synthesis and relevant algorithms, which rapidly expand the accessible chemical space to billions and trillions of compounds. (See under the section Synthesizability

for details) While substantial speed up can be achieved by brute-force hardware acceleration, such as computing on threads, clouds[193,194], and graphic processing units (GPU)[195,196], true scalability requires iterative approaches that prioritize likely binders for computational resources. In this section, we begin by exploring several filtering approaches that serve traditionally as the first-pass filters to reduce the chemical search space, then move on to active learning approaches that adaptively refine prioritization models based on feedback from previous docking results, and, finally, hierarchical approaches that explore divide-and-conquer-combine approaches to compound selection.

Filtering approaches

Filtering approaches are methods that eliminate unlikely hits based on molecular shape and pharmacophore complementarity[197] before more extensive evaluation. These methods are incorporated into programs, such as FTMap[198–200] and Pharmit[201,202]. In these programs, spatially arranged pharmacophores derived from ligands are aligned with landmarks ("druggable hotspots") modeled in the vacuum of a holo/apo target pocket to compute their volumetric correlation. The pharmacophores and landmarks are designed to share the same property fields, such as accounts of hydrophobic, aromatic, hydrogen-bonding, and charged interaction features. For instance, Pharmit leverages fragment docking to construct landmarks and the ligand candidates with matching pharmacophore features are then retrieved with specialized data structures. However, due to the reliance on volumetric overlaps, conformers of the ligand have to be substantially enumerated to cover bioactive conformations which may not be located at the relatively few unstrained minima[7,203]. Besides, feature engineering for landmarks/pharmacophore and fragment libraries can be complicated and highly target-dependent[204]. Recently, some progress has been made along these ideas. For instance, shape complementarity constraints were enforced in Kuzminynkh's variational autoencoder among other generative models[205–208]. The idea is to voxelate and blur the coordinates of atoms/pharmacophores/landmarks such that new molecules can be generated by interpreting the decoded voxels located in the proximity of the original coordinates. In Equivariant Scalar Field (ESF)[163], the scalar property fields are parametrized as several multi-channel equivariant graph neural networks[167], eliminating the reliance on handcrafted features.

Active learning approaches

Active learning approaches accelerate SBVS by iteratively selecting compounds for docking based on models trained on previously docked subsets. This process unfolds in successive rounds, aiming to bias future sampling toward high-affinity binders. For instance, Deep Docking[5,209] uses an initially random batch of docked ligands to iteratively generate training data for a ligand-based QSAR model. This CNN model then predicts docking score as "rewards" for the remaining compounds, prioritizing some of them for the next round of docking. In the CACHE Challenge #1, the Deep Docking approach screened 4.1 billion compounds from the Enamine REAL database against the LRRK2 WDR domain, yielding five validated hits (8.5% hit rate). In Molecular Pool-based Active Learning[210] (MolPAL), simple Bayesian optimization techniques, including Thompson Sampling (TS), were applied to guide the iterative selection of compounds via ligand-based ML/DL models. In TS, the expected reward of a compound is computed by sampling parameters on the posterior distribution ('belief'), which can be updated with new training data in a numerical closed form via conjugate distributions of the likelihood. Compounds are then prioritized by maximizing the expected reward. These strategies exemplify how simple integration of statistical inference techniques with traditional docking approaches can streamline the virtual screening process, expediting the identification of potential drug candidates.

Hierarchical approaches

In hierarchical approaches, molecular fragments are independently docked into a target and progressively elaborated into lead-like ligands for further evaluations only if the fragments demonstrate strong docking scores. This strategy is rooted in experimental and theoretical observations that the entropic loss in fragmental hit binding is compensated by enthalpically favorable interactions and that the binding modes of these fragmental substructures are frequently conserved in their corresponding lead-like compounds[211–213]. As a result, it helps to identify optimal substructures, which can be linked/grown/merged to form high-quality lead-like ligands. For instance, AutoGrow[214–216] is a genetic algorithm that iteratively elaborates a population of fragments with a symmetric vocabulary of functional groups, where conventional growing and merging mechanisms from fragment-based drug discovery campaigns[213,217] were integrated into its mutation and crossover operators, respectively. The objective of AutoGrow is to enhance docking scores across chemical species in multiple generations. Recently, generative methods, both ligand-based[218,219], and structure-based, such as PQR[220], have incorporated similar hierarchical strategies. Their general idea is to factorize the generative decision (e.g., for selection of the extending functional group and attachment points) as a product of conditional probability informed by ligand topology, target sequence, pocket conformation, and other features. Non-generative approaches, such as DeepFrag[221] and FRAME[222] have also gained traction. These methods aim to construct ligands in protein-ligand complexes, starting from fragments or hit-like ligands. In benchmarks involving PDB ligands with intentionally masked substructures, DeepFrag identified the correct extending functional group from a set of over 6500 options approximately 58% of the time; even when the exact lead-like ligand was not reproduced, the top predictions were often chemically similar, underscoring their potential as valid substitutions.

In another hierarchical modular approach, V-SYNTHES[11], a minimal fragment library (MEL) is prepared by enumerating synthons with corresponding reactions and capping groups. Upon docking into the receptor, the best MEL fragments are selected based on their docking scores and pocket volume available for enumeration at the capped attachment points. These other attachment points are then iteratively enumerated into a comprehensive product by selecting the best intermediates based on docking scores. As a result, the enumeration (~ 1 million compounds) is enriched with the best fragments, yielding high-scoring candidate hits. In the case of the cannabinoid CB2 receptor, the enrichment factor can yield about 250 and 450 against the random baseline for 2- and 3-component compounds, respectively. In a prospective screening study, using 11 billion compounds of Enamine REAL Space, V-SYNTHES achieved a 33% hit rate for the CB2 and a 28.5% hit rate for ROCK1 kinase, including many sub-micromolar hits. The performance for CB2 was much better than a standard non-hierarchical approach tested in parallel, which required nearly 100 times more computational resources. Notably, focused SAR-by-catalog of the best hits yielded compounds with nanomolar activity and high CB2/CB1 selectivity (200-fold), suggesting the utility of the method in the discovery of quality hits and lead series.

Synthesizability

Synthesizability is of utmost importance for any SBVS method to yield practical utility, because it determines whether the proposed molecules can be made and tested in the laboratory[223]. This concern is particularly pronounced in methods that generate molecules from scratch, rather than sampling from a pre-defined chemical library. Without restrictions on rules to produce these chemicals, these methods would lead to molecules that optimize computational qualities but are difficult or impossible to synthesize. (See Fig. 2) In practice, the synthesizability of predicted compounds is rarely confirmed through systematic experiments due to cost and scalability limitations. Instead, it is commonly assessed through expert opinion and retrosynthetic engines, which decide upon known tractable synthetic routes and the availability of building blocks. While synthesizability estimates may always have uncertainties arising from challenges in optimization of reaction conditions, or high cost of building blocks, they are pragmatic in selecting more feasible virtual hits. In this section, we will introduce the concept of synthesizability, several retrosynthetic engines, synthesizability proxies, and advances in incorporating considerations of synthesizability into computational programs.

Generative molecules and synthetic traps

Intuitively, chemical compounds can be generated one or several atoms at a time. However, while these algorithmically produced molecules may be chemically valid, hereby referred to as generative molecules, their chemical stability and synthetic feasibility are far from guaranteed, resulting in "synthetic traps". Several factors may contribute to this dilemma. First of all, the synthesis of real molecules, as complicated as those generative ones, is almost never performed linearly, but rather through multistep, convergent reactions. Hence, the "gap of making" between a generative molecule and its constituent synthons is rarely linear; molecules must be analyzed backward with simplifying transforms to mitigate synthetic difficulties. Secondly, even when large vendor-supplied building block libraries are available, these fragments are not uniformly interchangeable for arbitrary coupling reactions. Individual building blocks often differ in their chemical activation, stereoelectronic compatibility, and positional reactivity. A given attachment point on a synthon may be readily functionalized in the presence of one coupling partner yet remain inert or incompatible with others. This "mix-and-mismatch" in chemical behavior can easily render promising molecules synthetically inaccessible. Due to these asymmetries in reactivity, coupling synthons must often be divided into distinct, and sometimes mutually exclusive, sets of compatible synthons, each suited to specific reaction contexts (see asymmetric vocabulary, Fig. 2). For instance, many molecules generated by AutoGrow and other early hierarchical approaches[224–226] were not synthesizable. To address this, rule-based reaction restraints can be incorporated using a SMARTS-based reaction. SMARTS is a language for substructure pattern matching that enables the identification of reactive functional groups and reactive motifs among molecules. For example, AutoGrow3 and AutoGrow4 improved their mutation operators by incorporating 36 click-chemistry reactions from AutoClickChem[227] and 58 additional reaction templates from Hartenfeller et al.[228], thereby enforcing substructure-driven restraints (see shaded regions in Fig. 2b and d). While this improves synthesizability, SMARTS patterns may still overlook subtleties in reagent activation unless exhaustively defined, allowing hard-to-synthesize products to pass through. These limitations persist even in modern structure-based generative models that design ligands within binding pockets[20,229,230]. In many of these models, synthesizability is enforced by restraints, typically a numerical score or probability incorporated as a part of the loss function or sampling process. For instance, Synthetic Accessibility Score (SAScore)[231] is a widely adopted proxy to penalize counts of rare substructures from PubChem and synthetically complex molecular features, e.g., stereocenters, spiros, bridge heads and macrocycles. Other such proxies include SCScore and SMILES string length[231–233]. While such restraints can help steer optimization away from implausible molecules, they do not provide formal guarantees that the generated compounds are supported retrosynthetically when reassessed using a fully validated computer-aided synthesis planning (CASP) tool[234,235]. Moreover, they may fail to generalize as novel building blocks and synthetic methodologies evolve. Indeed, recent analyses of patented compounds in BindingDB reveal a statistical mode in SAScore distributions around a value of 3, with a long tail extending beyond 5, suggesting that many highly synthesizable ligands would not be even attempted if chemists rely on such rigid schemes as SAScore.

Computational tools to predict synthesizability

In principle, CASP tools can support the training and evaluation of generative molecules by providing on-the-fly assessments. These tools, such as Reaxys and Molecule. One, harness curated reaction rules and up-to-date inventories of purchasable precursors to assess synthesizability. However, they are generally impractical for real-time evaluation during training cycles due to their reliance on expert-defined rules, exhaustive substructure matching, and lack of acceleration. Recently, new tools have been developed to provide ground truths for proxy training or to act as post-generation validators[236–238]. For instance, AiZynthFinder[237] is an open-source tool based on Monte Carlo Tree Search (MCTS) guided by a neural network trained on reaction datasets, such as USPTO and Reaxys. The tool recursively decomposes target molecules into simpler precursors by applying reaction templates prioritized by a policy network, and optionally prunes implausible disconnections using a separate filter network. While AiZynthFinder achieves relatively fast retrosynthetic analysis (often within 10–60 s), its use of MCTS and multiple neural inferences per route still renders it computationally expensive for real-time model training. Another program ASKCOS[238], also uses template-based retrosynthesis models trained on large-scale reaction databases. However, in addition, it also integrates reaction condition recommendation, regioselectivity estimation, impurity prediction, and yield prediction. This delivers a chemically contextualized ecosystem encompassing both reverse (retrosynthetic) and forward (reaction outcome and condition) prediction. Recent work has sought to bypass the high computational cost of full retrosynthetic planning through such surrogates. For instance, RetroGNN[239] is a message-passing neural network trained to regress against the outputs of AiZynthFinder and Molecule.one. It achieves more than a 105-fold speedup while maintaining high fidelity (AUC > 0.99) in classifying hard-to-synthesize molecules, demonstrating the feasibility of integrating CASP-informed constraints directly into the training cycles of molecular generative models.

Table 3 | Examples of On-demand Ultra-large Chemical Spaces

Ultra-large Chemical Library	Year of Report	Number of Molecules	Availability	Reference
GSK XXL	2022	$1*10^{26}$	In-house	267
Pfizer GVL	2012	$1.3*10^{18}$	In-house	268
WuXi GalaXi	2022	$8 * 10^{9}$	Purchasable	241,269
Otava CHEMyria	2022	$1.2* 10^{10}$	Purchasable	269,270
Enamine REAL	2024	$6.5*10^{10}$	Purchasable	240,269
Enamine xREAL	2024	$2.4*10^{12}$	Purchasable	269,271
Freedom Space	2025	$1.4*10^{11}$	Purchasable	242

Constrained Generation In On-demand Chemical Spaces

Recent advances in parallel synthesis technologies and multicomponent reactions have enabled the design of ultra-large, on-demand chemical spaces, most prominently Enamine REAL[240] and WuXi GalaXi[241] (see Table 3 for other spaces). In contrast to traditional virtual screening libraries, which typically list available on-shelf molecules, the on-demand spaces are modularly constructed from a set of highly optimized reactions and combinatorial assemblies of validated synthons, which can be mapped to real-world building blocks that are filtered for reactivity, compatibility, and availability. The asymmetry in reactivity is conserved by grouping synthons in pre-compiled tables curated by experts or verified from CASP tools. This allows on-demand spaces to explore billions and trillions of compounds, with >80% success rates in synthesis, competitive prices, and delivery in 4-5 weeks, underscoring their utility in real-world drug discovery campaigns[5,11,242]. Multiple structure-based and ligand-based workflows have begun to explore these on-demand spaces using hierarchical enumeration or guided sampling. For example, platforms, such as V-SYNTHES[11] and Chemical Space Docking[243] () have demonstrated practical utility in prospective screening campaigns, leading to identification of new ligands chemotypes with nanomolar activities. Moreover, the Enamine REAL space[13] was adopted by CACHE organizers and participants to ensure that all proposed hits could be synthesized and tested experimentally. Other methods for chemical space exploration include TS (TS, see section Active Learning), e.g., in Klarich et al.[244], TS model evaluates each reagent in each reaction of the Enamine REAL space for likelihood of docking score associated with their enumerated product. This enabled the retrieval of 57–62 of the top 100 scoring molecules from a library of 335 million products by

sampling only 1% of the space, achieving a 100-fold speedup over brute-force docking.

Retrosynthesis-aware generative molecules

Back-testing with retrosynthesis is essential when generating novel compounds. Besides, in more advanced settings of SBVS, maintaining diversity across generative hits is critical. The sampling for a wider chemical space will fundamentally allow steering rooms for clinical properties and provide robust fallbacks against scoring biases. However, conventional generative models trained on limited, mostly unimodal datasets tend to converge on a small set of similar high-scoring compounds[245]. Even when synthetic accessibility heuristics are included, they lack strict enforcement, unlike the rigid constraints enforced by asymmetric vocabularies or retrosynthetic engines, leading to the generation of synthetically impractical molecules. Recently, these challenges have been tackled in SynFlowNet[246] by framing molecular generation as a generative flow network[247], sampling molecules proportionally to their reward and capturing multi-modal solutions in a reinforcement-learning setting. Its forward policy builds molecules stepwise using chemically validated reaction templates and purchasable building blocks. The backward policy in SynFlowNet does not directly evaluate synthetic feasibility but is instead trained to reconstruct valid backward trajectories that terminate at the initial building block state, enforcing retrosynthetic consistency within the directed acyclic graph of traversal. While not equivalent to explicit retrosynthesis engines, this procedure mirrors the logic of retrosynthetic analysis by ensuring that generated molecules can be decomposed into accessible precursors via valid reaction templates. The action space may also incorporate asymmetric vocabulary constraints to further restrict building-block choices, though the masking process can become verbose when reaction coverage is large. By modeling both forward and backward flows, SynFlowNet enforces that the stationary sampling distribution aligns with the reward function, which can include docking scores, synthetic-accessibility measures, or multi-parameter properties. This flow-based framework is particularly valuable for navigating multi-modal reward landscapes and adapting dynamically to changing synthesis constraints, such as evolving building-block availability due to supply shifts or new synthetic methods.

Summary and future directions

Recent developments in computational structure-based drug discovery have increasingly emphasized robust data management and standardized metadata to support machine learning applications for hit and lead discovery. Centralized, well-annotated databases and public–private partnerships have enhanced cross-lab reproducibility and facilitated data-driven model development[49]. Moreover, the scalability of ML-based methods has been significantly advanced through the use of knowledge-based filtering and hierarchical enumeration strategies, including the use of modular building blocks and ultra-large synthetically accessible spaces. Several machine-learning guided SBVS pipelines have now demonstrated experimental success, rivaling expensive high-throughput screening experimental workflows[248]. Unlike conventional high-throughput screening campaigns, constrained by physical stock availability, computational pipelines increasingly leverage on-demand synthesis of compound libraries, enabling vastly broader chemical exploration. Despite considerable progress in docking algorithms, scoring functions, and free energy estimation methods, the overall hit rates, as well as the quality and novelty of the hits in most prospective computational drug discovery campaigns, remain modest. While retrospective benchmarks often report favorable metrics, such as enrichment factors or RMSE values, these may not always translate to prospective screening outcomes. In virtual screening, specifically, the vast excess of non-binders over binders in random chemical libraries (often exceeding 10000:1) means that an enrichment factor at 1% (EF1%) well above 100 is required merely to achieve a probability of at least one true binder among 100 experimentally tested molecules. By contrast, most benchmarking studies report EF1% values in the modest 20-40 range, revealing a substantial shortfall between the theoretical performance needed for realistic hit discovery. The success of computational pipelines in practical settings is still constrained by the limited accuracy of scoring functions and affinity estimates, resulting in the dominance of false positives, as well as challenges related to protein flexibility. Moreover, even high-confidence in silico predictions can misfire in the real-life discovery cycle due to neglected factors, such as the target conformational/functional states, synthetic tractability, or even assay artifacts. As a result, the true utility of the computational programs depends not only on algorithmic sophistication but also on rigorous system setup, careful interpretation of model outputs, and close integration with medicinal chemistry and biology workflows.

At the same time, synergistic integration of physics-based and knowledge-based methods still remains underdeveloped. Physics-based approaches are principled and grounded in physical laws. They encode symmetry, conservation principles, and functional forms that reflect decades of validated theoretical insight across physics, chemistry, and chemical physics. However, these methods are often computationally expensive due to their reliance on explicit sampling (e.g., molecular dynamics, free energy perturbation). They frequently require well-informed initial guesses or ansatzes to accelerate convergence. This presents a natural opportunity for knowledge-based methods, which excel at generating educated guesses and priors. Their strength lies in speed and statistical bounds on uncertainty when properly calibrated. More fundamentally, we envision three principled forms of synergism between physics and machine learning.

First, the incorporation of physical inductive biases into machine learning models, as demonstrated in the framework of physics-informed neural networks (PINNs)[249]. The essence of this approach is to incorporate the analytical structure of physical laws, typically expressed as differential equations or closed-form interactions, into equations of the machine learning models. This family of methods can include NNPs that build upon foundational frameworks like the ACE, which mirrors the multipole expansion of many-body interactions. These models benefit from analytically tractable and interpretable forms and can often outperform black-box deep learning models when data are sparse or when extrapolation is needed. Notably, hybrid models on electronic structures have also emerged where missing pieces of theory, such as the exchange-correlation (XC) functional in DFT, are approximated using machine learning, leaving other parts of theory unaltered[250,251]. Semi-empirical methods[104,252–254], such as extended tight-binding[255,256] based on dencoupled clustersity functional tight-binding theory[257] are also rapidly changing the landscape of accurate energy predictions.

Second, computed physical quantities (from physics-based theories unaltered) can be directly injected as features. These features, beyond simple atomic coordinate geometries, such as electronic density projections[129,258], electrostatic potential surface descriptors[170] and QM topological descriptors[259] are capturing feasible short- and long-range interaction patterns akin to qualitative/empirical pharmacophore descriptors. A prominent example is the use of Quantum Theory of Atoms in Molecules (QTAIM), where critical points in the electron density distribution are extracted. These features systematically characterize atomic and bond-level properties, such as electron density, Laplacian, ellipticity, and localization functions, all of which have been shown to correlate with interaction strength and binding affinity in various QSAR studies[260–262]. They can serve as features to inform machine learning models meaningfully.

Third, aligning sampling strategies in machine learning with enhanced sampling techniques from physics is becoming increasingly pivotal[263]. Knowledge-based methods, such as generative models, excel at proposing educated guesses, which are crucial for targeting relevant physical basins within the high-dimensional conformational landscape. By initializing simulations near these basins, the overall sampling can become more efficient, accelerating convergence toward thermodynamically significant regions. Examples presented in RAVE[264,265] and BioEmu[266] are efficiently uncovering rare events (conformation transition states, ligand dissociation) through proposals of low-dimensional latent coordinates, principled on variational bounds. This exemplifies how ML-derived priors can guide physics-based sampling, yielding near-optimal exploration of potential energy surfaces.

Looking ahead, major challenges remain in scaling machine-learned models outside their training regimes and integrating them seamlessly into laboratory processes, such as synthesis design and binding validation. However, the principled combination of statistical inference and physical theory, augmented by expert chemical knowledge, offers a promising path forward. Realizing this integration will require interdisciplinary collaboration to bridge algorithmic advances with experimental realities.

Data availability

No datasets were generated or analyzed during the current study.

Received: 8 March 2025; Accepted: 1 November 2025;
Published online: 01 December 2025

References

1. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
2. Tropsha, A., Isayev, O., Varnek, A., Schneider, G. & Cherkasov, A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nat. Rev. Drug Discov.* **23**, 141–155 (2024).
3. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
4. 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).
5. Gentile, F. et al. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat. Protoc.* **17**, 672–697 (2022).
6. 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).
7. Gu, S., Smith, M. S., Yang, Y., Irwin, J. J. & Shoichet, B. K. Ligand Strain Energy in Large Library Docking. *J. Chem. Inf. Model.* **61**, 4331–4341 (2021).
8. Brooijmans, N. & Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annu. Rev. Biophys* **32**, 335–373 (2003).
9. Huang, S.-Y., Grinter, S. Z. & Zou, X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **12**, 12899–12908 (2010).
10. Muegge, I. & Rarey, M. Small Molecule Docking and Scoring. in *Reviews in Computational Chemistry* 1–60 (Wiley, 2001). https://doi.org/10.1002/0471224413.ch1.
11. Sadybekov, A. A. et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022).
12. Gorgulla, C. et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).
13. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
14. Unke, O. T. et al. Machine Learning Force Fields. *Chem. Rev.* **121**, 10142–10186 (2021).
15. Kovács, D. P., Batatia, I., Arany, E. S. & Csányi, G. Evaluation of the MACE force field architecture: from medicinal chemistry to materials science. *J. Chem. Phys.* **159**, 044118 (2023).
16. Pantaleão, S. Q., Fernandes, P. O., Gonçalves, J. E., Maltarollo, V. G. & Honorio, K. M. Recent advances in the prediction of pharmacokinetics properties in drug design studies: a review. *ChemMedChem* **17**, e202100542 (2022).
17. Pushpakom, S. et al. Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2019).
18. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol. Pharm.* **14**, 3098–3104 (2017).
19. Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G. & Boström, J. Deep reinforcement learning for multiparameter optimization in de novo drug design. *J. Chem. Inf. Model.* **59**, 3166–3176 (2019).
20. Loeffler, H. H. et al. Reinvent 4: modern AI–driven generative molecule design. *J. Cheminformatics* **16**, 20 (2024).
21. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
22. Wallach, I. & Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **58**, 916–932 (2018).
23. Sieg, J., Flachsenberg, F. & Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **59**, 947–961 (2019).
24. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res* **28**, 235–242 (2000).
25. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
26. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
27. Blass, B. E. Chapter 1 - Drug Discovery and Development: An Overview of Modern Methods and Principles. in *Basic Principles of Drug Discovery and Development* (ed. Blass, B. E.) 1–34 (Academic Press, Boston, 2015). https://doi.org/10.1016/B978-0-12-411508-8.00001-3.
28. 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).
29. Harder, E. et al. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **12**, 281–296 (2016).
30. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).
31. Katritch, V., Totrov, M. & Abagyan, R. ICFF: A new method to incorporate implicit flexibility into an internal coordinate force field. *J. Comput. Chem.* **24**, 254–265 (2003).
32. Setiadi, J. et al. Tuning potential functions to host–guest binding data. *J. Chem. Theory Comput.* **20**, 239–252 (2024).
33. Nelson, L. et al. Implementation of the QUBE force field in somd for high-throughput alchemical free-energy calculations. *J. Chem. Inf. Model.* **61**, 2124–2130 (2021).
34. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein–ligand docking using GOLD. *Proteins Struct. Funct. Bioinforma.* **52**, 609–623 (2003).
35. Ross, G. A. et al. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Commun. Chem.* **6**, 1–12 (2023).
36. Moore, J. H. et al. Automated relative binding free energy calculations from SMILES to ΔΔG. *Commun. Chem.* **6**, 1–12 (2023).
37. Bieniek, M. K. et al. An open-source molecular builder and free energy preparation workflow. *Commun. Chem.* **5**, 1–9 (2022).
38. Zdrazil, B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* **52**, D1180–D1192 (2024).
39. Bento, A. P. et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* **42**, D1083–D1090 (2014).
40. Thomas, P. D. & Dill, K. A. Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457–469 (1996).
41. Buttenschoen, M., Morris, G. M. & Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci* **15**, 3130–3139 (2024).

42. Durairaj, J. et al. PLINDER: the protein-ligand interactions dataset and evaluation resource. bioRxiv https://doi.org/10.1101/2024.07.17.603955 (2024).
43. Kramer, C. et al. The Need for Continuing Blinded Pose- and Activity Prediction Benchmarks. *J. Chem. Inf. Model*. https://doi.org/10.1021/acs.jcim.4c02296 (2025).
44. Kramer, C. & Gedeck, P. Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *J. Chem. Inf. Model.* **50**, 1961–1969 (2010).
45. Francoeur, P. G. et al. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *J. Chem. Inf. Model.* **60**, 4200–4215 (2020).
46. Yan, J., Friedrich, S. & Kurgan, L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief. Bioinform.* **17**, 88–105 (2016).
47. Tang, G. W. & Altman, R. B. Knowledge-based fragment binding prediction. *PLOS Comput. Biol.* **10**, e1003589 (2014).
48. Su, M. et al. Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).
49. Edfeldt, K. et al. A data science roadmap for open science organizations engaged in early-stage drug discovery. *Nat. Commun.* **15**, 5640 (2024).
50. ten Brink, T. & Exner, T. E. Influence of protonation, tautomeric, and stereoisomeric states on protein−ligand docking results. *J. Chem. Inf. Model.* **49**, 1535–1546 (2009).
51. Cho, A. E., Guallar, V., Berne, B. J. & Friesner, R. Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem.* **26**, 915–931 (2005).
52. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1. *J. Mol. Biol.* **285**, 1735–1747 (1999).
53. Hartmann, C., Antes, I. & Lengauer, T. Docking and scoring with alternative side-chain conformations. *Proteins Struct. Funct. Bioinforma.* **74**, 712–726 (2009).
54. Davis, A. M., St-Gallay, S. A. & Kleywegt, G. J. Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov. Today* **13**, 831–841 (2008).
55. Pintilie, G. et al. Measurement of atom resolvability in cryo-EM maps with Q-scores. *Nat. Methods* **17**, 328–334 (2020).
56. Barad, B. A. et al. EMRinger: side chain–directed model and map validation for 3D cryo-electron microscopy. *Nat. Methods* **12**, 943–946 (2015).
57. Lawson, C. L. et al. Outcomes of the EMDataResource cryo-EM Ligand Modeling Challenge. *Nat. Methods* **21**, 1340–1348 (2024).
58. Bender, B. J. et al. A practical guide to large-scale docking. *Nat. Protoc.* **16**, 4799–4832 (2021).
59. Nakamura, T., Wang, X., Terashi, G. & Kihara, D. DAQ-Score Database: assessment of map–model compatibility for protein structure models from cryo-EM maps. *Nat. Methods* **20**, 775–776 (2023).
60. Kufareva, I., Ilatovskiy, A. V. & Abagyan, R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res* **40**, D535–D540 (2012).
61. Kim, S. et al. PubChem 2023 update. *Nucleic Acids Res* **51**, D1373–D1380 (2023).
62. Liu, T. et al. BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Res*. gkae1075 https://doi.org/10.1093/nar/gkae1075 (2024).
63. Hunter, F. M. I. et al. Drug and Clinical Candidate Drug Data in ChEMBL. *J. Med. Chem*. https://doi.org/10.1021/acs.jmedchem.5c00920 (2025).
64. Papadatos, G. et al. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res* **44**, D1220–D1228 (2016).
65. Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The experimental uncertainty of heterogeneous public ki data. *J. Med. Chem.* **55**, 5165–5173 (2012).
66. Landrum, G. A. & Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **64**, 1560–1567 (2024).
67. Heyndrickx, W. et al. MELLODDY: cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information. *J. Chem. Inf. Model.* **64**, 2331–2344 (2024).
68. Edwards, A. M. & Owen, D. R. Protein–ligand data at scale to support machine learning. *Nat. Rev. Chem.* **9**, 634–645 (2025).
69. Kunnakkattu, I. R. et al. PDBe CCDUtils: an RDKit-based toolkit for handling and analysing small molecules in the Protein Data Bank. *J. Cheminformatics* **15**, 117 (2023).
70. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **48**, 4111–4119 (2005).
71. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
72. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
73. Su, M., Feng, G., Liu, Z., Li, Y. & Wang, R. Tapping on the black box: how is the scoring power of a machine-learning scoring function dependent on the training set? *J. Chem. Inf. Model.* **60**, 1122–1136 (2020).
74. Li, H. et al. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinformatics* **35**, 3989–3995 (2019).
75. Lam, J. H. et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat. Commun.* **10**, 4941 (2019).
76. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
77. Škrinjar, P., Eberhardt, J., Tauriello, G., Schwede, T. & Durairaj, J. Have protein-ligand cofolding methods moved beyond memorisation? Preprint at https://doi.org/10.1101/2025.02.03.636309 (2025).
78. Liu, F. et al. The impact of library size and scale of testing on virtual screening. *Nat. Chem. Biol*. 1–7 https://doi.org/10.1038/s41589-024-01797-w (2025).
79. Huang, N., Shoichet, B. K. & Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **49**, 6789–6801 (2006).
80. Stein, R. M. et al. Property-unmatched decoys in docking benchmarks. *J. Chem. Inf. Model.* **61**, 699–714 (2021).
81. Mysinger, M. M., Carchia, M., Irwin, J. ohnJ. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
82. Irwin, J. J. & Shoichet, B. K. ZINC − A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).
83. McGann, M. FRED and HYBRID docking performance on standardized datasets. *J. Comput. Aided Mol. Des.* **26**, 897–906 (2012).
84. Neves, M. A. C., 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).
85. Repasky, M. P. et al. Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. *J. Comput. Aided Mol. Des.* **26**, 787–799 (2012).

86. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
87. Imrie, F., Bradley, A. R., van der Schaar, M. & Deane, C. M. Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *J. Chem. Inf. Model.* **58**, 2319–2330 (2018).
88. Bottegoni, G., Kufareva, I., Totrov, M. & Abagyan, R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. *J. Med. Chem.* **52**, 397–406 (2009).
89. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
90. Lee, A., Lee, K. & Kim, D. Using reverse docking for target identification and its applications for drug discovery. *Expert Opin. Drug Discov* **11**, 707–715 (2016).
91. Brown, B. P., Mendenhall, J., Geanes, A. R. & Meiler, J. General purpose structure-based drug discovery neural network score functions with human-interpretable pharmacophore maps. *J. Chem. Inf. Model.* **61**, 603–620 (2021).
92. 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).
93. Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **49**, 1079–1093 (2009).
94. Li, Y., Han, L., Liu, Z. & Wang, R. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *J. Chem. Inf. Model.* **54**, 1717–1736 (2014).
95. Kufareva, I., Katritch, V., Stevens, R. C. & Abagyan, R. Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. *Structure* **22**, 1120–1139 (2014).
96. Ferrara, P., Gohlke, H., Price, D. J., Klebe, G. & Brooks, C. L. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **47**, 3032–3047 (2004).
97. Wang, R., Lu, Y. & Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **46**, 2287–2303 (2003).
98. Herasymenko, O. et al. CACHE challenge #2: targeting the RNA site of the SARS-CoV-2 Helicase Nsp13. *J. Chem. Inf. Model.* **65**, 6884–6898 (2025).
99. Mobley, D. L. & Dill, K. A. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure* **17**, 489–498 (2009).
100. Mark, A. E. & van Gunsteren, W. F. Decomposition of the free energy of a system in terms of specific interactions: implications for theoretical and experimental studies. *J. Mol. Biol.* **240**, 167–176 (1994).
101. Smith, P. E. & van Gunsteren, W. F. When are free energy components meaningful? *J. Phys. Chem.* **98**, 13735–13740 (1994).
102. Mardirossian, N. & Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **115**, 2315–2372 (2017).
103. Hobza, P. Calculations on noncovalent interactions and databases of benchmark interaction energies. *Acc. Chem. Res.* **45**, 663–672 (2012).
104. Christensen, A. S., Kubař, T., Cui, Q. & Elstner, M. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Rev.* **116**, 5301–5337 (2016).
105. Greengard, L. & Rokhlin, V. A new version of the fast multipole method for the laplace equation in three dimensions. *Acta Numer* **6**, 229–269 (1997).
106. Barnes, J. & Hut, P. A hierarchical O(N log N) force-calculation algorithm. *Nature* **324**, 446–449 (1986).
107. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
108. Vanommeslaeghe, K., Raman, E. P. & MacKerell, A. D. Jr. Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **52**, 3155–3168 (2012).
109. Roos, K. et al. OPLS3e: extending force field coverage for drug-like small molecules. *J. Chem. Theory Comput.* **15**, 1863–1874 (2019).
110. Sellers, B. D., James, N. C. & Gobbi, A. A comparison of quantum and molecular mechanical methods to estimate strain energy in druglike fragments. *J. Chem. Inf. Model.* **57**, 1265–1275 (2017).
111. Devereux, C. et al. Extending the Applicability of the ANI deep learning molecular potential to sulfur and halogens. *J. Chem. Theory Comput.* **16**, 4192–4202 (2020).
112. Wang, L.-P., Chen, J. & Van Voorhis, T. Systematic parametrization of polarizable force fields from quantum chemistry data. *J. Chem. Theory Comput.* **9**, 452–460 (2013).
113. Wang, L.-P., Martinez, T. J. & Pande, V. S. Building force fields: an automatic, systematic, and reproducible approach. *J. Phys. Chem. Lett.* **5**, 1885–1891 (2014).
114. Jing, Z. et al. Polarizable force fields for biomolecular simulations: recent advances and applications. *Annu. Rev. Biophys.* **48**, 371–394 (2019).
115. Bissantz, C., Kuhn, B. & Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **53**, 5061–5084 (2010).
116. Kuhn, B., Gilberg, E., Taylor, R., Cole, J. & Korb, O. How significant are unusual protein–ligand interactions? insights from database mining. *J. Med. Chem.* **62**, 10441–10455 (2019).
117. Freitas, R. F. de & Schapira, M. A systematic analysis of atomic protein–ligand interactions in the PDB. *MedChemComm* **8**, 1970–1981 (2017).
118. Mackerell, A. D. Jr, Feig, M. & Brooks III, C. L. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25**, 1400–1415 (2004).
119. Haghiri, S., Viquez Rojas, C., Bhat, S., Isayev, O. & Slipchenko, L. ANI/EFP: modeling long-range interactions in ani neural network with effective fragment potentials. *J. Chem. Theory Comput.* **20**, 9138–9147 (2024).
120. Grisafi, A. & Ceriotti, M. Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **151**, 204105 (2019).
121. Stern, C. D. et al. Capturing non-local through-bond effects in molecular mechanics force fields: II. Using fractional bond orders to fit torsion parameters. Preprint at https://doi.org/10.1101/2022.01.17.476653 (2022).
122. Arnautova, Y. A., Abagyan, R. A. & Totrov, M. Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling. *Proteins Struct. Funct. Bioinforma.* **79**, 477–498 (2011).
123. Arnautova, Y. A., Abagyan, R. & Totrov, M. All-atom internal coordinate mechanics (ICM) force field for hexopyranoses and glycoproteins. *J. Chem. Theory Comput.* **11**, 2167–2186 (2015).
124. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
125. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process. Syst.* **35**, 11423–11436 (2022).
126. Zhao, Q., Anstine, D. M., Isayev, O. & Savoie, B. M. Δ2 machine learning for reaction property prediction. *Chem. Sci.* **14**, 13392–13401 (2023).

127. Zubatyuk, R., Smith, J. S., Nebgen, B. T., Tretiak, S. & Isayev, O. Teaching a neural network to attach and detach electrons from molecules. *Nat. Commun.* **12**, 4870 (2021).
128. Allen, A. E. A. et al. Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning. *npj Comput. Mater.* **10**, 1–9 (2024).
129. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
130. Anstine, D.M., Zubatyuk, R. & Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chem. Sci.* **16**, 10228–10244 (2025).
131. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
132. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
133. Akimov, A. V. & Prezhdo, O. V. Large-scale computations in chemistry: a bird's eye view of a vibrant field. *Chem. Rev.* **115**, 5797–5890 (2015).
134. Raghavan, B. et al. Drug design in the exascale era: a perspective from massively parallel QM/MM simulations. *J. Chem. Inf. Model.* **63**, 3647–3658 (2023).
135. Shen, L., Wu, J. & Yang, W. Multiscale quantum mechanics/molecular mechanics simulations with neural networks. *J. Chem. Theory Comput.* **12**, 4934–4946 (2016).
136. Böselt, L., Thürlemann, M. & Riniker, S. Machine learning in QM/MM molecular dynamics simulations of condensed-phase systems. *J. Chem. Theory Comput.* **17**, 2641–2658 (2021).
137. Semelak, J. A. et al. Advancing multiscale molecular modeling with machine learning-derived electrostatics. *J. Chem. Theory Comput.* **21**, 5194–5207 (2024).
138. Giudetti, G. et al. How reproducible are QM/MM simulations? Lessons from computational studies of the covalent inhibition of the SARS-CoV-2 main protease by carmofur. *J. Chem. Theory Comput.* **18**, 5056–5067 (2022).
139. Zwanzig, R. W. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **22**, 1420–1426 (1954).
140. Kollman, P. eter Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.* **93**, 2395–2417 (1993).
141. Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **3**, 300–313 (1935).
142. Christ, C. D., Mark, A. E. & van Gunsteren, W. F. Basic ingredients of free energy calculations: a review. *J. Comput. Chem.* **31**, 1569–1582 (2010).
143. Blazhynska, M., Goulard Coderc de Lacam, E., Chen, H. & Chipot, C. Improving speed and affordability without compromising accuracy: standard binding free-energy calculations using an enhanced sampling algorithm, multiple-time stepping, and hydrogen mass repartitioning. *J. Chem. Theory Comput.* **19**, 3091–3101 (2023).
144. Miller, B. R. I. et al. MMPBSA.py: an efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* **8**, 3314–3321 (2012).
145. Jorgensen, W. L. & Thomas, L. L. Perspective on free-energy perturbation calculations for chemical equilibria. *J. Chem. Theory Comput.* **4**, 869–876 (2008).
146. Aldeghi, M., Bluck, J. P. & Biggin, P. C. Absolute alchemical free energy calculations for ligand binding: a beginner's guide. in *Computational Drug Discovery and Design* 199–232 (Humana Press, New York, NY, 2018). https://doi.org/10.1007/978-1-4939-7756-7_11.
147. Aldeghi, M., Heifetz, A., Bodkin, M. J., Knapp, S. & Biggin, P. C. Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.* **7**, 207–218 (2015).
148. Cournia, Z., Allen, B. & Sherman, W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J. Chem. Inf. Model*. https://doi.org/10.1021/acs.jcim.7b00564 (2017).
149. Gapsys, V. et al. Accurate absolute free energies for ligand–protein binding based on non-equilibrium approaches. *Commun. Chem.* **4**, 1–13 (2021).
150. Ohadi, D. et al. Input pose is key to performance of free energy perturbation: benchmarking with monoacylglycerol lipase. *J. Chem. Inf. Model*. https://doi.org/10.1021/acs.jcim.4c01223 (2024).
151. Wang, L. et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **137**, 2695–2703 (2015).
152. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288 (1982).
153. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking1. *J. Mol. Biol.* **267**, 727–748 (1997).
154. 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).
155. Abagyan, R., Totrov, M. & Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488–506 (1994).
156. Goodsell, D. S. & Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins Struct. Funct. Bioinforma.* **8**, 195–202 (1990).
157. Davis, I. W. & Baker, D. RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol.* **385**, 381–392 (2009).
158. Di Nola, A., Roccatano, D. & Berendsen, H. J. C. Molecular dynamics simulation of the docking of substrates to proteins. *Proteins Struct. Funct. Bioinforma.* **19**, 174–182 (1994).
159. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129 (1990).
160. Blum, C. & Roli, A. Hybrid Metaheuristics: An Introduction. in *Hybrid Metaheuristics: An Emerging Approach to Optimization* (eds Blum, C., Aguilera, M. J. B., Roli, A. & Sampels, M.) 1–30 (Springer, 2008). https://doi.org/10.1007/978-3-540-78295-7_1.
161. Abagyan, R. & Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002 (1994).
162. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. S. DiffDock: diffusion steps, twists, and turns for molecular docking. In *Proceedings of The Eleventh International Conference on Learning Representations* (2023).
163. Jing, B., Jaakkola, T. S. & Berger, B. Equivariant scalar fields for molecular docking with fast fourier transforms. NeurIPS 2023 AI for Science Workshop (2023).
164. McNutt, A. T. et al. GNINA 1.0: molecular docking with deep learning. *J. Cheminformatics* **13**, 43 (2021).
165. Dunn, I. et al. CACHE Challenge #1: docking with GNINA Is All You Need. *J. Chem. Inf. Model*. https://doi.org/10.1021/acs.jcim.4c01429 (2024).
166. Raush, E., Abagyan, R. & Totrov, M. Graph-convolutional neural net model of the statistical torsion profiles for small organic molecules. *J. Chem. Inf. Model.* **62**, 5896–5906 (2022).
167. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
168. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, D. R. & Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. in *Proc. 39th International Conference on Machine Learning* 20503–20521 (PMLR, 2022).

169. Cao, D. et al. SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction. *Nat. Methods* **22**, 310–322 (2025).
170. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
171. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, D439–D444 (2022).
172. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
173. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
174. Karelina, M., Noh, J. J. & Dror, R. O. How accurately can one predict drug binding modes using AlphaFold models? *eLife* **12**, 89386 (2023).
175. Lyu, J. et al. AlphaFold2 structures guide prospective ligand discovery. *Science* **384**, eadn6354 (2024).
176. Kaynak, B. T., Bahar, I. & Doruker, P. Essential site scanning analysis: a new approach for detecting sites that modulate the dispersion of protein global motions. *Comput. Struct. Biotechnol. J.* **18**, 1577–1586 (2020).
177. Bahar, I., Lezon, T. R., Yang, L.-W. & Eyal, E. Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* **39**, 23–42 (2010).
178. Stank, A., Kokh, D. B., Fuller, J. C. & Wade, R. C. Protein binding pocket dynamics. *Acc. Chem. Res.* **49**, 809–815 (2016).
179. Katritch, V., Rueda, M. & Abagyan, R. Ligand-guided receptor optimization. *Methods Mol. Biol. Clifton NJ* **857**, 189–205 (2012).
180. Calderón, J. C., Ibrahim, P., Gobbo, D., Gervasio, F. L. & Clark, T. General metadynamics protocol to simulate activation/deactivation of class A GPCRs: proof of principle for the serotonin receptor. *J. Chem. Inf. Model.* **63**, 3105–3117 (2023).
181. Porter, J. R. et al. Cooperative changes in solvent exposure identify cryptic pockets, switches, and allosteric coupling. *Biophys. J.* **116**, 818–830 (2019).
182. Vithani, N. et al. Exploration of cryptic pockets using enhanced sampling along normal modes: a case study of KRAS G12D. *J. Chem. Inf. Model.* **64**, 8258–8273 (2024).
183. Zimmerman, M. I. & Bowman, G. R. FAST conformational searches by balancing exploration/exploitation trade-offs. *J. Chem. Theory Comput.* **11**, 5747–5757 (2015).
184. Gur, M., Zomot, E., Cheng, M. H. & Bahar, I. Energy landscape of LeuT from molecular simulations. *J. Chem. Phys.* **143**, 243134 (2015).
185. Bryant, P., Kelkar, A., Guljas, A., Clementi, C. & Noé, F. Structure prediction of protein-ligand complexes from sequence information with Umol. *Nat. Commun.* **15**, 4536 (2024).
186. Qiao, Z., Nie, W., Vahdat, A., Miller, T. F. & Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nat. Mach. Intell.* **6**, 195–208 (2024).
187. Passaro, S. et al. Boltz-2: towards accurate and efficient binding affinity prediction. Preprint at https://doi.org/10.1101/2025.06.14.659707 (2025).
188. Masters, M. R., Mahmoud, A. H. & Lill, M. A. Investigating whether deep learning models for co-folding learn the physics of protein ligand interactions. *Nat. Commun.* **16**, 8854 (2024).
189. Volochnyuk, D. M. et al. Evolution of commercially available compounds for HTS. *Drug Discov. Today* **24**, 390–402 (2019).
190. Irwin, J. J. et al. ZINC20—A free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
191. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
192. Tingle, B. I. et al. ZINC-22—A free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model.* **63**, 1166–1176 (2023).
193. Tingle, B. I. & Irwin, J. J. Large-scale docking in the cloud. *J. Chem. Inf. Model.* **63**, 2735–2741 (2023).
194. Grebner, C. et al. Virtual screening in the cloud: how big is big enough? *J. Chem. Inf. Model.* **60**, 4274–4282 (2020).
195. Santos-Martins, D. et al. Accelerating AutoDock4 with GPUs and gradient-based local search. *J. Chem. Theory Comput.* **17**, 1060–1073 (2021).
196. Yu, Y. et al. Uni-Dock: GPU-accelerated docking enables ultralarge virtual screening. *J. Chem. Theory Comput.* **19**, 3336–3345 (2023).
197. Nicholls, A. et al. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.* **53**, 3862–3886 (2010).
198. Katchalski-Katzir, E. et al. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA* **89**, 2195–2199 (1992).
199. Kozakov, D. et al. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **10**, 733–755 (2015).
200. Padhorny, D. et al. Protein–protein docking by fast generalized Fourier transforms on 5D rotational manifolds. *Proc. Natl. Acad. Sci. USA* **113**, E4286–E4293 (2016).
201. Sunseri, J. & Koes, D. R. Pharmit: interactive exploration of chemical space. *Nucleic Acids Res* **44**, W442–W448 (2016).
202. Koes, D. R. & Camacho, C. J. Shape-based virtual screening with volumetric aligned molecular shapes. *J. Comput. Chem.* **35**, 1824–1834 (2014).
203. Tong, J. & Zhao, S. Large-scale analysis of bioactive ligand conformational strain energy by ab initio calculation. *J. Chem. Inf. Model.* **61**, 1180–1192 (2021).
204. Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discov. Today* **15**, 444–450 (2010).
205. Imrie, F., Hadfield, T. E., Bradley, A. R. & Deane, C. M. Deep generative design with 3D pharmacophoric constraints. *Chem. Sci.* **12**, 14577–14589 (2021).
206. Imrie, F., Bradley, A. R., van der Schaar, M. & Deane, C. M. Deep generative models for 3D linker design. *J. Chem. Inf. Model.* **60**, 1983–1995 (2020).
207. Ragoza, M., Masuda, T. & Ryan Koes, D. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem. Sci.* **13**, 2701–2713 (2022).
208. Kuzminykh, D. et al. 3D molecular representations based on the wave transform for convolutional neural networks. *Mol. Pharm.* **15**, 4378–4385 (2018).
209. Gentile, F. et al. Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent. Sci.* **6**, 939–949 (2020).
210. 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).
211. Jencks, W. P. On the attribution and additivity of binding energies. *Proc. Natl. Acad. Sci.* **78**, 4046–4050 (1981).
212. Page, M. I. & Jencks, W. P. Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proc. Natl. Acad. Sci. USA* **68**, 1678–1683 (1971).
213. Murray, C. W. & Rees, D. C. The rise of fragment-based drug discovery. *Nat. Chem.* **1**, 187–192 (2009).
214. Durrant, J. D., Amaro, R. E. & McCammon, J. A. AutoGrow: a novel algorithm for protein inhibitor design. *Chem. Biol. Drug Des.* **73**, 168–178 (2009).
215. Durrant, J. D., Lindert, S. & McCammon, J. A. AutoGrow 3.0: an improved algorithm for chemically tractable, semi-automated protein inhibitor design. *J. Mol. Graph. Model.* **44**, 104–112 (2013).
216. Spiegel, J. O. & Durrant, J. D. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. *J. Cheminformatics* **12**, 25 (2020).

217. Rees, D. C., Congreve, M., Murray, C. W. & Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discov.* **3**, 660–672 (2004).
218. Jin, W., Barzilay, R. & Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. in *Proc. 35th International Conference on Machine Learning* 2323–2332 (PMLR, 2018).
219. Pedawi, A., Gniewek, P., Chang, C., Anderson, B. & van den Bedem, H. An efficient graph generative model for navigating ultra-large combinatorial synthesis libraries. *Adv. Neural Inf. Process. Syst.* **35**, 8731–8745 (2022).
220. Chan, L., Kumar, R., Verdonk, M. & Poelking, C. A multilevel generative framework with hierarchical self-contrasting for bias control and transparency in structure-based ligand design. *Nat. Mach. Intell.* **4**, 1130–1142 (2022).
221. Green, H. & Durrant, J. D. DeepFrag: an open-source browser app for deep-learning lead optimization. *J. Chem. Inf. Model.* **61**, 2523–2529 (2021).
222. Powers, A. S. et al. Geometric deep learning for structure-based ligand design. *ACS Cent. Sci.* **9**, 2257–2267 (2023).
223. Stanley, M. & Segler, M. Fake it until you make it? Generative de novo design and virtual screening of synthesizable molecules. *Curr. Opin. Struct. Biol.* **82**, 102658 (2023).
224. 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).
225. Gillet, V., Johnson, A. P., Mata, P., Sike, S. & Williams, P. SPROUT: a program for structure generation. *J. Comput. Aided Mol. Des.* **7**, 127–153 (1993).
226. Rotstein, S. H. & Murcko, M. A. GroupBuild: a fragment-based method for de novo drug design. *J. Med. Chem.* **36**, 1700–1710 (1993).
227. Durrant, J. D. & McCammon, J. A. AutoClickChem: click chemistry in silico. *PLOS Comput. Biol.* **8**, e1002397 (2012).
228. Hartenfeller, M. et al. A collection of robust organic synthesis reactions for in silico molecule design. *J. Chem. Inf. Model.* **51**, 3093–3098 (2011).
229. Luo, S., Guan, J., Ma, J. & Peng, J. A 3D generative model for structure-based drug design. *Adv. Neural Inf. Process. Syst.* **34**, 6229–6239 (2021).
230. Igashov, I. et al. Equivariant 3D-conditional diffusion model for molecular linker design. *Nat. Mach. Intell.* **6**, 417–427 (2024).
231. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminformatics* **1**, 8 (2009).
232. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
233. Voršilák, M., Kolář, M., Čmelo, I. & Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminformatics* **12**, 35 (2020).
234. Gao, W. & Coley, C. W. The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.* **60**, 5714–5723 (2020).
235. Renz, P., Van Rompaey, D., Wegner, J. K., Hochreiter, S. & Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discov. Today Technol* **32–33**, 55–63 (2019).
236. Rethinking Molecule Synthesizability with Chain-of-Reaction. https://arxiv.org/html/2509.16084v1.
237. Saigiridharan, L. et al. AiZynthFinder 4.0: developments based on learnings from 3 years of industrial application. *J. Cheminformatics* **16**, 57 (2024).
238. Tu, Z. et al. ASKCOS: open-source, data-driven synthesis planning. *Acc. Chem. Res.* **58**, 1764–1775 (2025).
239. Liu, C.-H. et al. RetroGNN: fast estimation of synthesizability for virtual screening and de novo design by learning from slow retrosynthesis software. *J. Chem. Inf. Model.* **62**, 2293–2300 (2022).
240. REAL Space - Enamine. https://enamine.net/compound-collections/real-compounds/real-space-navigator.
241. WuXi GalaXi. https://www.labnetwork.com/frontend-app/p/#!/library/virtual.
242. Kapeliukha, A. *et al*. Freedom Space 3.0: ML-Assisted Selection of Synthetically Accessible Small Molecules. *J. Chem. Inf. Model*. https://doi.org/10.1021/acs.jcim.5c01912 (2025).
243. Beroza, P. et al. Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. *Nat. Commun.* **13**, 6447 (2022).
244. Klarich, K., Goldman, B., Kramer, T., Riley, P. & Walters, W. P. Thompson Sampling—An Efficient Method for Searching Ultralarge Synthesis on Demand Databases. *J. Chem. Inf. Model.* **64**, 1158–1171 (2024).
245. Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R. & Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Comput. Mol. Sci.* **12**, e1608 (2022).
246. Cretu, M. et al. SynFlowNet: Design of Diverse and Novel Molecules with Synthesis Constraints. In *Proc. The Thirteenth International Conference on Learning Representations* (ed. Cretu, M. et al.) (2025).
247. Bengio, Y. et al. GFlowNet Foundations. *J. Mach. Learn. Res.* **24**, 1–55 (2023).
248. Xu, Z. et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nat. Med.* **31**, 2602–2610 (2025).
249. Raissi, M. Deep hidden physics models: deep learning of nonlinear partial differential equations. *J. Mach. Learn. Res.* **19**, 1–24 (2018).
250. Kirkpatrick, J. et al. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* **374**, 1385–1389 (2021).
251. Kulik, H. J. et al. Roadmap on machine learning in electronic structure. *Electron. Struct.* **4**, 023004 (2022).
252. Nováček, M. & Řezáč, J. PM6-ML: the synergy of semiempirical quantum chemistry and machine learning transformed into a practical computational method. *J. Chem. Theory Comput.* **21**, 678–690 (2025).
253. Pecina, A., Fanfrlík, J., Lepšík, M. & Řezáč, J. SQM2.20: semiempirical quantum-mechanical scoring function yields DFT-quality protein–ligand binding affinity predictions in minutes. *Nat. Commun.* **15**, 1127 (2024).
254. Pecina, A. et al. SQM/COSMO Scoring Function at the DFTB3-D3H4 Level: Unique Identification of Native Protein–Ligand Poses. *J. Chem. Inf. Model.* **57**, 127–132 (2017).
255. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
256. Froitzheim, T., Müller, M., Hansen, A. & Grimme, S. g-xTB: A General-Purpose Extended Tight-Binding Electronic Structure Method For the Elements H to Lr (Z=1–103). Preprint at https://doi.org/10.26434/chemrxiv-2025-bjxvt (2025).
257. Elstner, M. et al. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **58**, 7260–7268 (1998).
258. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
259. Matta, C. F. & Boyd, R. J. An Introduction to the Quantum Theory of Atoms in Molecules. in *The Quantum Theory of Atoms in Molecules* 1–34 (Wiley, 2007). https://doi.org/10.1002/9783527610709.ch1.
260. Isert, C., Atz, K., Riniker, S. & Schneider, G. Exploring protein–ligand binding affinity prediction with electron density-based geometric deep learning. https://doi.org/10.1039/D3RA08650J (2024).
261. Karelson, M., Lobanov, V. S. & Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **96**, 1027–1044 (1996).
262. Ginex, T., Vázquez, J., Estarellas, C. & Luque, F. J. Quantum mechanical-based strategies in drug discovery: finding the pace to

new challenges in drug design. *Curr. Opin. Struct. Biol.* **87**, 102870 (2024).
263. Cersonsky, R. K., Cheng, B., De Vivo, M. & Tiwary, P. Machine learning and statistical mechanics: shared synergies for next generation of chemical theory and computation. *J. Chem. Theory Comput.* **21**, 5359–5364 (2025).
264. Mehdi, S., Smith, Z., Herron, L., Zou, Z. & Tiwary, P. Enhanced sampling with machine learning. *Annu. Rev. Phys. Chem.* **75**, 347–370 (2024).
265. Ribeiro, J. M. L., Bravo, P., Wang, Y. & Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J. Chem. Phys.* **149**, 072301 (2018).
266. Lewis, S. et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* **0**, eadv9817 (2025).
267. 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).
268. 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).
269. 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).
270. 12 Billion Novel Molecules: CHEMriya - OTAVA's On-Demand Chemical Space. https://www.otavachemicals.com/products/chemriya.
271. infiniSee xREAL • Your Access Point to Trillions of Drug-Like Compounds. *BioSolveIT* https://www.biosolveit.de/products/infinisee_xreal/.

Acknowledgements

J.H.L. acknowledges funding from the Croucher Foundation of Hong Kong. V.K. acknowledges National Institutes of Health grant R35GM153437. The authors thank Michael K. Gilson, Renxiao Wang, Gabriele Corso and Alexander MacKerell for fruitful discussions.

Author contributions

J.H.L. and V.K. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Jordy Homing Lam or Vsevolod Katritch.

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025