npj | drug discovery

Perspective

https://doi.org/10.1038/s44386-024-00001-2

# Structure and dynamics in drug discovery



Haixin Wei ✉ & J. Andrew McCammon ✉

The roles of computing in structure-based drug discovery are considered, from early studies based on some of the first experimental structures of enzyme-inhibitor complexes, through the use of advanced molecular dynamics simulations and machine learning methods. This perspective aims to explore the history, current trends, and future directions of these methodologies.

Developing new drugs is very expensive and time-consuming[1,2]. Generally speaking, the drug discovery and development process usually takes around 10–14 years and more than 1 billion dollars[3]. In the 1980s, the use of computers was extended from data handling to a more prominent role in drug discovery[4], and since then, computational approaches in the drug discovery, design, and development process have been adopted rapidly[2,5–10]. It has been estimated that by the use of computer-aided drug discovery (CADD) approaches, the cost of drug discovery and development can be reduced up to 50%[11].

CADD is a specialized discipline that uses computational methods to simulate drug-receptor interactions to determine if a given molecule will bind to a target, and if so, what its affinity would be[12]. There are mainly two types of CADD techniques: ligand-based drug design (LBDD) and structure-based drug design (SBDD). SBDD can only be used when the three-dimensional structures of target proteins are known, while LBDD design is employed in situations that the structures are unavailable. In recent years, there have been simultaneous advancements in structural biology such as cryo-EM, and computational protein structure prediction as with AlphaFold[13], allowing for access to many more high-resolution 3D structures of novel drug-receptor complexes[14–17]. Because of the convergence of these breakthroughs, SBDD, as a structurally enabled computational method (Fig. 1), has become a driving force for the discovery of novel small molecule therapeutics.

Once a well-defined drug target structure is obtained, virtual screening of libraries of compounds is conducted, through molecular docking software. By scoring and ranking a collection of ligands, potential drug candidates are identified. Thus, the rapid expansion of drug-like chemical space, easily accessible for hit and lead discovery, is another key factor driving the advancement of CADD. Just a few years ago, the compound libraries were still limited to several million on-shelf compounds from vendors and in-house screening libraries in pharma. Now, screening can be done with ultra-large virtual libraries and chemical spaces of drug-like compounds, which can be readily made on-demand, rapidly growing beyond billions of compounds[18].

However, there are certain limitations of SBDD despite its many successes. One of the biggest remaining challenges in SBDD is target flexibility. Proteins and ligand molecules possess high flexibility in solution and undergo frequent conformational changes. However, most molecular docking tools allow for high flexibility of the ligand, but the protein is kept fixed or provided with only limited flexibility to the residues present within or near the active site. It is very difficult to provide complete molecular flexibility to the protein as this increases the space and time complexity of the computation dramatically. The second challenge is the difficulty of exploring cryptic pockets, which are not shown in the original structure but will be revealed by protein conformation changes. Those pockets often relate to allosteric regulations, which would offer extra opportunities targeting beyond the primary endogenous binding site of the receptor.

One solution to the above challenges is molecular dynamics simulation, also referred as MD simulation[19]. Nowadays, MD simulation has become one of the most influential methods for modeling conformational changes within a ligand-target complex upon binding a small molecule[20,21]. Nevertheless, it is commonly believed that normal MD is unable to cross substantial energy barriers within a simulation's lifespan, preventing it from efficiently traversing the energy landscape of a protein in complex with ligands. By adding a boost potential to smooth the system potential energy surface, accelerated molecular dynamics[22,23] (aMD) methods were developed for helping with this issue. The boost potential decreases the energy barriers and therefore accelerates transitions between the different low-energy states[24]. With this, aMD is able to sample distinct biomolecular conformations and help with the receptor flexibility and cryptic pockets problems.

In a word, the discovery of new therapeutic drugs has been greatly speeded by computer-aided methods based on molecular structure and dynamics. This perspective provides a brief appraisal of the origin, role, and prospects of these methods, starting from early work based on x-ray crystallographic structures of proteins, through the use of molecular dynamics simulations, to early work drawing upon simulation and machine learning.

## Structure-based drug discovery

### Increase in available structures

Perhaps the earliest use of protein structures for drug discovery was that for the development of the important inhibitors of angiotensin-converting enzyme (ACE), captopril, and enalopril. The design of these drugs, used to treat high blood pressure and other conditions, benefitted from modeling based on the crystallographic structure of carboxypeptidase A, which has a similar active site featuring a catalytically important zinc ion[25,26].

Departments of Chemistry & Biochemistry, and Pharmacology, University of California San Diego, La Jolla, CA, USA.
✉e-mail: h9wei@ucsd.edu; jmccammon@ucsd.edu

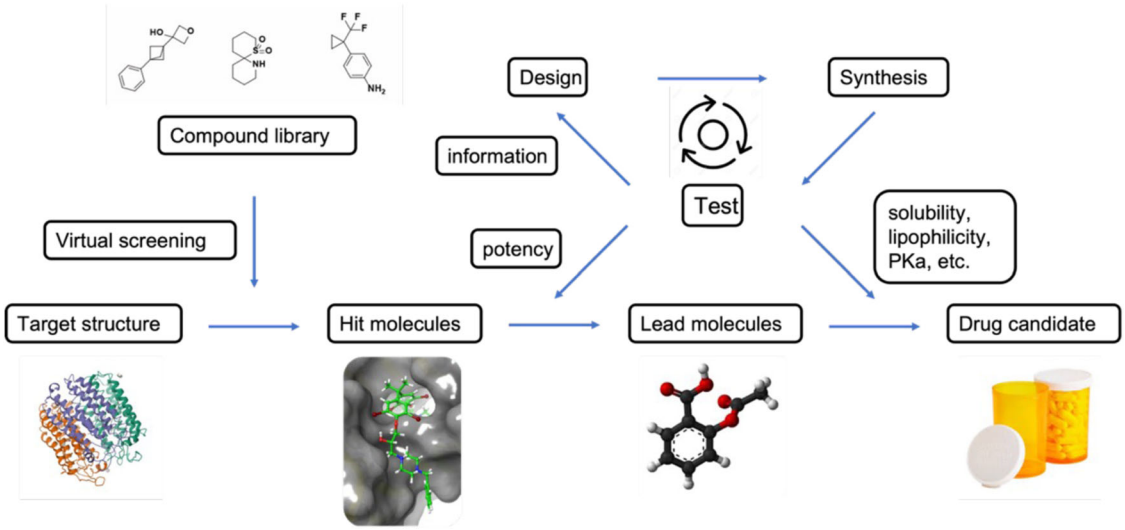
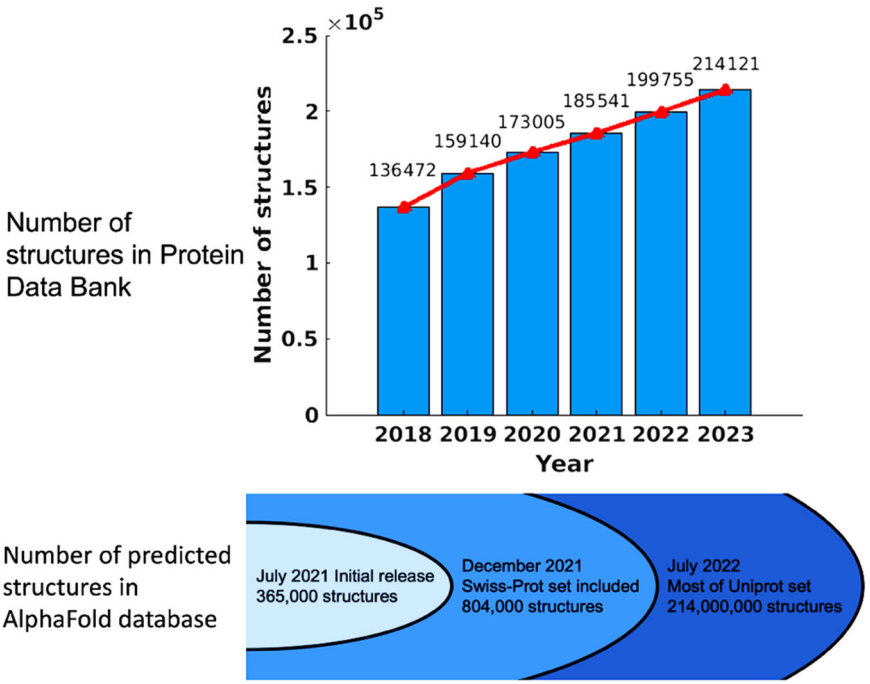



**Fig. 1 | Schematic of structure-based drug discovery.** The computational methods are usually employed in two stages of SBDD. The virtual screening is widely adapted in the initial hit molecule searching, and many predictive models are used in the design-synthesis-test cycle.

**Fig. 2 | The increasing of available structures.** The rapid expansion of drug target structures in both the Protein Data Bank and AlphaFold database, which has significantly increased opportunities for discovering new drugs.




This field has grown rapidly in the ensuing years, due in part to the elucidation of the structures of thousands of proteins, nucleic acids, and other potential drug targets[27]. In recent years, due to leaps in structural biology, including automation in crystallography[15], microcrystallography[14], and cryo-electron microscopy technology[16,17,28], the 3D structures for many clinically important targets have been revealed, often in a state relevant to its biological function (Fig. 2). Especially impressive has been the recent structural revolution for G protein-coupled receptors (GPCRs)[29], ion channels[30,31] and other membrane proteins that mediate the action of more than half of drugs[32], providing excellent targets for ligand screening and lead optimization.

The number of target structures has increased significantly with the arrival of machine learning tools such as AlphaFold, which reliably predict the atomic structure of proteins for which experimental structures may not be available[13]. Since its launch in 2021, the AlphaFold Protein Structure Database has had released over 214 million unique protein structures[33], compared to around 200,000 PDB structures corresponding to approximately 60,000 unique protein sequences (Fig. 2). This new data set almost covers the complete UniProt database. Additionally, AlphaFold models can cover the entire length of protein sequences compared to the fragmented, often short coverage of PDB entries. Clearly, researchers and pharmaceutical companies can try structure-based drug discovery techniques using these models, presenting unprecedented opportunities for targets without a prior experimental structure.

## Improving computational methods

The early years of structure-based drug discovery typically featured computational studies of the binding of small molecules to active sites of

enzymes, using structure-based virtual screening or molecular docking, with the test molecules chosen from libraries of compounds or analogs of known binders[34]. The quality of binding of each test molecule was typically determined using model potential energy functions chosen to balance speed and accuracy[34].

Docking molecules of a virtual drug-like compound library into a target receptor structure and predicting its binding score is a major step in a structure-based drug discovery campaign, which plays a key role in any successful application[7,29,35]. The predicted candidate ligand sets, produced by such virtual screening, usually show useful hit rates, about 10%-40% in experimental testing[36]. Some novel hits may also exhibit noteworthy potencies, in the 0.1–10-μM range, for different types of targets[36].

Special attention has been devoted to ligand scoring functions, which are supposed to reliably select top binders and to rule out false-positive predictions. This is especially important with the growth of library size. For example, a one-in-a-million rate of false positives in a billion-compound library would result in a thousand false hits, which obviously complicates the selection of hit candidates. Another major challenge is the computation cost. With increasing library sizes, the computational time of docking itself is the main bottleneck in virtual screening processes. Nowadays, screenings on ultra-large virtual libraries that include billions of drug-like compounds are feasible, thanks to the recent availability of cloud computing and graphics processing unit (GPU) computing resources[36].

### Expansion of accessible chemical space

A successful structure-based drug discovery screening campaign depends critically on diverse ligand libraries that cover a large part of the chemical space that might be of interest. There are at least two reasons why the compound library should be large and diverse. Firstly, such a library will obviously increase the chance of identifying potential hits in the virtual screening process[37]. This has been demonstrated in many ultra-large virtual screening campaigns, for example in refs. 38–42. Secondly, the diversity of drug candidates will be improved, offering more opportunities for further optimization and modification. A good library can expand the chemical diversity, novelty, and patentability of the hits[43]. The hit analogs in the same library can help build a meaningful structure-activity relationship, which further facilitates downstream optimization steps.

Several approaches have been developed recently to boost the size and diversity of screening libraries, among which the virtual on-demand libraries are most worth mentioning. In 2017, the readily accessible (REAL) database by Enamine[18] was established and became the first commercially available on-demand library. The REAL library uses carefully selected in-stock building blocks and optimized parallel synthesis protocols, which makes it a fast, reliable source of compounds[39]. Through years of development, the fully enumerated REAL database has grown from approximately 170 million compounds in 2017 to more than 6.7 billion compounds in 2024. The successful application of the REAL database has been recently documented in several virtual drug screening campaigns[44], some of which showed exquisite performance with nanomolar and even sub-nanomolar affinities. Some other similar ultra-large virtual libraries are also available, for example, synthetically accessible virtual inventory (SAVI)[45] developed by the US National Institutes of Health.

## Dynamics-based drug discovery

### The relaxed complex method

Structure-based screening and docking methods are often able to sample accessible conformations of the small molecules being tested but, apart from modest relaxation, they are typically limited in the extent to which they can sample different conformations of the target molecule. Related to this, the docking methods are limited in their ability to estimate free energies of binding.

The introduction of molecular dynamics (MD) simulations for the study of proteins opened the way to improved methods for drug discovery[46,47]. MD simulations allow sampling not only of the conformations of the ligands, but also those of the target molecule. In the target molecule, pre-existing pockets for potential binding vary somewhat in size and shape during its normal dynamics. Importantly, cryptic pockets may also appear, providing new binding sites that might be accessed in docking and modeling. A systematic approach to representing this variation in potential binding sites is the Relaxed Complex Method (Fig. 3), in which representative target conformations, often including novel, cryptic binding sites, are selected from MD simulations for use in docking studies[48,49].

An early example of the effective use of MD in drug discovery is the development of the first FDA-approved inhibitor of HIV integrase. MD simulations[50] starting with x-ray crystallographic structures of the core domain of this integrase, determined by the Davies group[51], provided early indications of significant flexibility in the active site region. The Davies group subsequently reported the first crystal structure of an inhibitor-bound form of the protein[52]. The Relaxed Complex Method was subsequently applied to explore the binding of this inhibitor to different conformations of the protein generated by MD[53]. Surprisingly, the results showed that the binding pocket seen in the crystallographic structure expanded to yield a trench in the protein surface that could accommodate two different binding poses of the inhibitor, rotated by about 180 degrees and translationally separated[53]. This discovery was noted by researchers at Merck & Co[54], who pursued the two-pose concept in developing the first FDA-approved inhibitor of the integrase for clinical use[55].

### The enhanced sampling methods

Classic MD often shows a debilitating timescale problem, since its integration timestep is restricted to a few femtoseconds. As such, a large number of enhanced sampling methods have been proposed over the decades, to accelerate the dynamics of the system and access much longer timescales[56–58]. These methods can be broadly classified into two categories.

The first category includes methods based on collective variables. These methods, such as umbrella sampling[59], metadynamics[57], and the weighted ensemble method[60], utilize collective variables to represent the degrees of freedom of interest, thereby reducing the dimensionality of the free energy surface and enabling more comprehensive sampling of these regions. However, selecting effective collective variables can be challenging and often

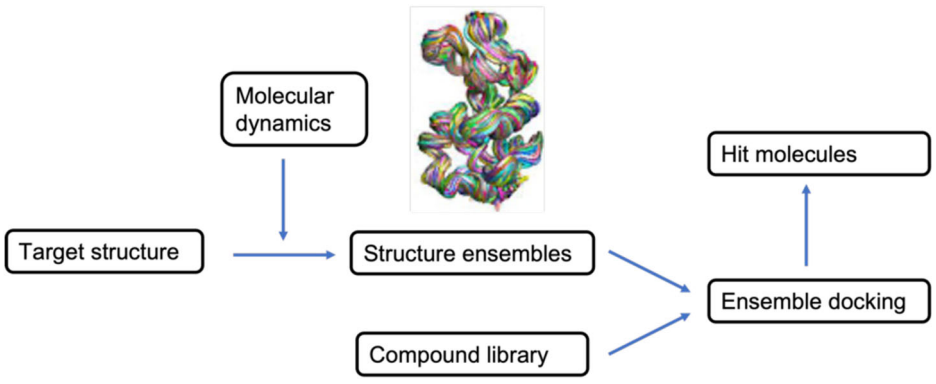



**Fig. 3 | The workflow of dynamic-based drug discovery.** The dynamic-based drug discovery usually uses structure ensembles that are generated by molecular dynamics or enhanced sampling to enrich the diversity of hit molecules.

requires prior knowledge of energy basins and their potential exiting pathways. As a result, the applicability of such methods is somewhat constrained. For instance, weighted ensemble methods are particularly effective for sampling transitions between well-defined end states.

The second category encompasses methods that do not rely on collective variables and comprises a variety of distinct approaches. Replica exchange dynamics, for example, employs multiple copies of the same simulation at different temperatures to enhance the probability of observing rare events[61]. The mixed Monte Carlo–MD method[56,62] combines Monte Carlo simulations with molecular dynamics to explore extended timescales. Additionally, accelerated MD methods have gained increasing attention in recent drug discovery research, for their advantages in simulating target conformations and identifying cryptic binding sites.

Accelerated MD methods[22,23] are enhanced sampling approaches that work by adding a nonnegative boost potential to smooth the system potential energy surface. No collective variables are employed, so that the results can be directly reweighted to yield the probability of occurrence of conformations and other thermodynamic properties[38]. This allows the accelerated MD methods to be applied in most situations. A more recent Gaussian Accelerated Molecular Dynamics[63,64] (GaMD) method has been developed in which the boost potential follows a Gaussian distribution and allows for simple and accurate reweighting of the simulations.

Employing these accelerated MD simulation techniques has at least two benefits: better accounting for the target flexibility and increasing the likelihood of finding cryptic binding sites. In Miao, et al.[65], the M2 muscarinic acetylcholine receptor (mAChR) was used as a GPCR model, and accelerated MD simulations were used to account for the receptor flexibility. Through iterative molecular docking and experimental testing, the authors successfully identified positive and negative allosteric modulators of M2 mAChR, with unprecedented chemical diversity and, remarkably, demonstrated in vivo selectivity of the targeted receptor in a family of similar receptors. In Seitz, etc.[66], through GaMD, computational screening and in vitro activity testing, the authors reported novel scaffolds for inhibiting *Mycobacterium tuberculosis* cytochrome bd oxidase, again validated experimentally.

## Free energy calculations

An important advantage in the use of MD in drug discovery is that it opens the way to the computation of free energies of binding. The components of free energy of a system, such as enthalpy and entropy, reflect averages over its allowed structures. Free energy studies are a natural next step after library screening and docking studies, providing a way to rank promising lead compounds. Computation of relative free energies of binding, which allows for estimation of the relative affinities of congeners or other pairs of molecules, can help to guide screening or synthetic work. Such calculations can be done using the computational alchemy formalism, first demonstrated for proteins in the binding of different ligands to the enzyme trypsin and a mutant form of the enzyme[67,68]. Standard free energies of binding of individual ligands is also possible, as demonstrated in 1988[69] and then in a more complete form in 1997[70].

### Application in hit optimization

Iterative design-make-test cycles are the essential way to turn an initial hit molecule into a lead drug candidate, which requires properties that are aligned with relevant clinical and pharmaceutical demands. Eventually, the developed drug will need to balance a group of properties like drug efficacy, efficiency, toxicity, human tolerance, etc. The early hit-to-lead stage is focused primarily on binding affinity because high-affinity compounds are likely to be active at lower doses and to have longer residence times in the relevant receptors. Much progress has been made in the past decades for rapidly optimizing binding affinity, most notably by utilizing free energy perturbation (FEP) calculations[71–81]. Previously, it has been challenging for free energy calculations to achieve the accuracy, efficiency, and reliability required for application to the hit optimization process. Fortunately, recent advances in methodologies and the development of specialized parallel computing devices (GPUs), have enabled free energy calculations to attain a level that is sufficient to drive better synthesis decisions[82]. Such free energy calculations, combined with molecular dynamic simulations, can be used to examine molecular conformations, molecular motion, intermolecular interactions, etc. With this level of accuracy and insight, medicinal chemists can test their ideas based on the FEP calculations before actual synthesis and reduce the percentage of compounds synthesized that do not meet potency requirements.

### Application in initial screening

Due to its high computational costs (compared with docking), free energy calculations are not usually applied in the initial screening phase. However, recently both industrial and academic groups are embracing binding free energy calculations as a screening tool, as featured by a few prospective applications. The advantage of free energy calculation is that it is more accurate than the conventional docking process, yet more rapid than what can be done with experiments, which allows drug discovery teams to explore a larger part of chemical space. There has been a growing literature, as we show below, demonstrating the value of free-energy simulations in early drug discovery applications.

In 2016, Merck KGaA started a large initiative to prospectively assess the prediction accuracy of relative binding free energy calculations. In this large-scale assessment concerning 12 targets and 23 chemical series which are from their active drug discovery projects, more than 35,000 free energy perturbation calculations were done, over 400 novel molecules with predicted high binding affinities were identified and subsequently synthesized and tested[83]. Another example came from Bayer, where a team used a computationally empowered workflow to identify novel covalent allosteric binders for the KRAS G12C isoform[84]. In one of the largest free-energy simulation studies on a GPCR, a collection of 3,4- dihydropyrimidine-2(1H)-ones were enumerated and computationally screened with free-energy simulations against the A2B adenosine receptor[85]. In another study of a viral protein, free-energy simulations were performed on the crystal structure of norovirus RNA-dependent RNA polymerase in a complex with several known binders to determine binding free energies of these molecules relative to the natural nucleotide substrates, through which a virtual nucleotide library containing 121 molecules was screened and two novel molecules were successfully identified with in vitro activity[86].

## Machine learning

Machine learning is a promising new tool for drug discovery. Perhaps the major contribution of machine learning in structure-based drug discovery to date has been through the dramatic expansion of available target structures. In the context of structure and dynamics methods for ligand-receptor recognition, AlphaFold, and related methods are already providing input for protein modeling studies utilizing MD[87].

### AlphaFold

One of the most dramatic developments in recent years in structural biology is the appearance of AlphaFold[13]. To the date of this review, DeepMind released the updated version, AlphaFold 3[88]. Accurate target structures are critical to the mission of structure-based drug discovery, as the starting point of any campaign. Since the introduction of the first AlphaFold model[13], a revolution in modeling the structure of proteins has started. Enormous progress has been achieved in protein structure prediction with the development of AlphaFold 1, and the field has grown tremendously with a number of later methods that build on the ideas and techniques of AlphaFold 2[33,89–96]. Clearly, the emergence of AlphaFold models opens unprecedented opportunities for drug discovery with targets that were not available before.

AlphaFold 3 is capable of predicting the joint structure of complexes including proteins, nucleic acids, small molecules, ions, and modified residues. This is promising for structure-based drug discovery, as it is commonly believed that a complex structure with both target and a known binder is a better starting point than a single apo structure. The authors of

AlphaFold 3 claim in their publication that the performance of their models on protein-ligand systems is better than classical docking tools, such as Vina[97,98], and greatly outperforms all other blind dockings like RoseTTA-Fold All-Atom. Their evaluation was done on their PoseBusters benchmark set, which is composed of 428 protein-ligand structures that were not included in their training. The reported accuracy, as the percentage of protein-ligand pairs with pocket-aligned ligand root mean squared deviation of less than 2 Å, is over 90% for their high-confidence group.

AlphaFold and other Machine learning models may be limited in the near future in their ability to predict cryptic pockets, given limitations in training data. It may be possible to create training data by enhanced sampling MD and use this data to train machine learning models to produce additional target structures. Related research is in Lyu, etc.[99], where the authors followed up on several hundred computational hits and found that there was little to no overlap for the same receptor when starting with the AlphaFold 2 model versus the experimental structure. This indicated that AlphaFold models are already showing some potential on modeling different conformations.

## Machine learning docking scoring

One of the attractive directions for applying machine-learning techniques in structure-based drug discovery is the generation of machine-learned docking score functions, by extracting information from both known ligand activities and corresponding protein-ligand 3D structures. The time seems ripe for such models as the databases needed have been established, for example, the PDBbind database[100]. Thus, there have already been some attempts in this direction, with various approaches to represent the data and different network architectures, including spatial graph-convolutional models[101,102], 3D deep convolutional neural networks[103,104] or their combinations[105]. However, most of these models have suffered severe overtraining problems[106]. One possible explanation for this phenomenon is that the PDBbind database does not have an appropriate number of negative cases, i.e., ligands with non-optimal binding interfaces to enrich the training set. One solution to this could be to train the model with the results of physics-based docking, in addition to experimental data. A recent example of such a model is RTCNN[107], although its practical utility remains to be demonstrated.

Since the scoring functions, whether physics-based or machine-learned, are primarily designed or trained to effectively separate potential binders from non-binders, it is probably the best idea to combine both of them and wield their maximum advantages in different situations. If the physics-based and machine-learned scoring functions are relatively independent and emphasize different interactions between the target and ligands, their combination can perhaps reduce the false-positive rates and improve the quality of the hits. For example, in Durrant and McCammon[108], they describe a fast and accurate neural-network-based scoring function that can be used to re-score the docked poses of candidate ligands. Another such example is in the latest 3DR Grand Challenge 4 results for ligand IC50 predictions[109], in which the top methods that used a combination of both physics-based and machine-learned scoring outperformed others. Further, for more accurate potency predictions, the smaller focused library of candidate binders selected by the initial machine learning methods can be sent to more elaborate physics-based analysis, such as free energy perturbation methods[83,110–112]. There is no doubt, with its striking capabilities, machine learning will play an important role in virtual screening campaigns, as well as in lead optimization stages.

## Other machine-learning models

Early machine learning algorithms such as support vector machine, and random forest have been used to predict ligand activities and properties, even before deep learning/neural networks came into wider use. For example, quantitative structure-property relationship (QSPR) models which were used to predict pharmacodynamic and pharmacokinetic properties, such as solubility, lipophilicity, etc., have been trained based on large and broad experimental datasets[113–115]. Machine learning techniques are also implemented for many quantitative structure-activity relationship (QSAR) algorithms[116], where the training set and the resulting models will focus on a specific target and one type of chemical scaffold. These machine learning models can greatly help on guiding the hit-to-lead and lead optimization process. Other machine learning models aimed at drug repurposing have also been suggested, based on extensive ligand-target binding datasets, chemical similarity clustering, etc.[117,118].

For QSPR, large public and private databases have been accumulated, with various properties such as solubility, lipophilicity or in vitro proxies for oral bioavailability and brain permeability experimentally measured for many thousands of diverse compounds, allowing prediction of these properties in a broad range of new compounds. However, things are not quite similar for the situation of QSAR models, because, for different target classes, the data availability vary greatly. Probably, the most abundant QSAR databases are those for the kinase superfamily and aminergic GPCRs, as they are popular as drug targets. Recently, an unbiased benchmark of the best machine learning QSAR models was given by a IDG-DREAM Drug-Kinase Binding Prediction Challenge, with the best models achieving a root-mean-square error of 0.95 for the predicted versus experimental $pK_d$ values[119]. Such accuracy may be acceptable in screenings for the initial hits for less explored kinases and guiding lead optimization. However, as noted before, the kinase family is unique as it is the largest target family with more than 500 structures. The performance and generalizability of such machine learning models for other less rich target families remain unclear.

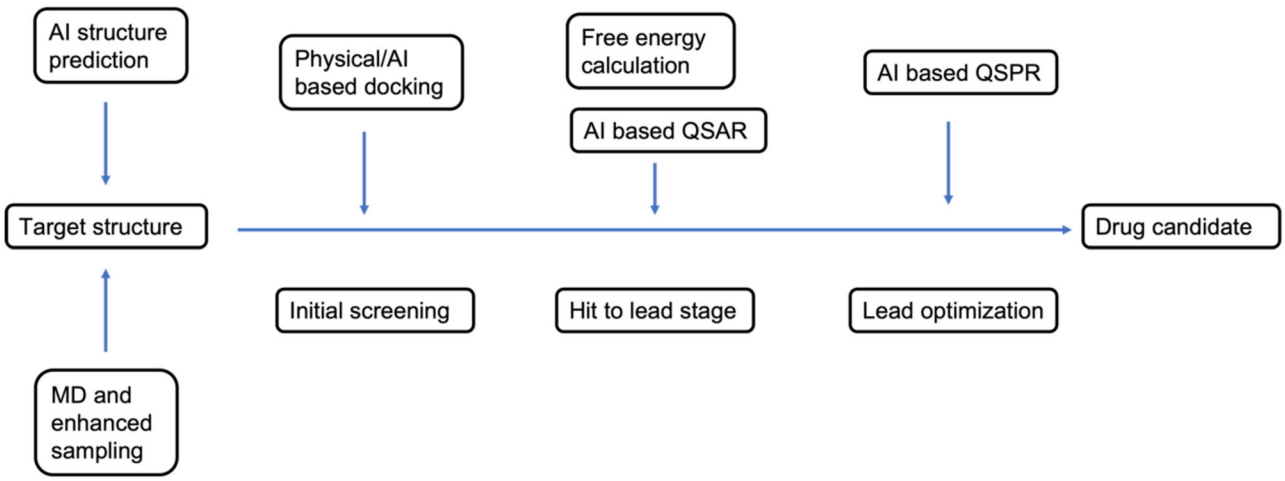



**Fig. 4 | Paradigm of future computer-driven drug discovery.** Comparing with the standard computer-aided drug discovery pipeline, the computer-driven drug discovery relies primarily on computational tools.

## Future directions

In the last decades, with all the successes and challenges, the transformation from computer-aided drug discovery to computer-driven drug discovery is emerging. The rapid advancements in structural biology along with the bloom of computational protein-structure prediction, are allowing for access to many more high-resolution 3D structures of novel drug-receptor complexes. Dynamic-based drug discovery strategies, coupled with advanced sampling methods, are offering new possibilities for drug design. During the early hit identification stage, the ultra-scale virtual screening approaches, both structure-based and AI-based, are becoming ready in providing fast and cost-effective start points into drug discovery campaigns. The rapid expansion of accessible chemical space, together with the dramatic increases in computational power such as GPUs and cloud computing, are resulting in the ability to virtually screen multi-billions of drug-like chemical space. At the hit-to-lead stage, the more elaborate potency prediction tools such as free energy perturbation and AI-based QSAR often guide rational optimization of ligand potency. In the end, data-driven computational tools are used for multi-parameter optimization, including solubility, permeability, and pharmacokinetic properties, to identify the final drug candidate. It is evident that these breakthroughs are converging towards a new era of computer-driven drug discovery (Fig. 4).

## Data availability

No datasets were generated or analysed during the current study.

Received: 19 July 2024; Accepted: 25 September 2024;
Published online: 07 November 2024

## References

1. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the last 25 years. *J. Nat. Prod.* **70**, 461–477 (2007).
2. Office, C. B. *Research and development in the pharmaceutical industry: a CBO study*. (Government Printing Office, 2007).
3. Daina, A. et al. Drug design workshop: A web-based educational tool to introduce computer-aided drug design to the general public. *J. Chem. Educ.* **94**, 335–344 (2017).
4. Kaul, P.N. (1998). Drug discovery: Past, present and future. In: Jucker, E. (eds) Progress in Drug Research. Progress in Drug Research, vol 50. Birkhäuser, Basel. https://doi.org/10.1007/978-3-0348-8833-2_1
5. Van Drie, J. H. Computer-aided drug design: the next 20 years. *J. Comput.-Aided Mol. Des.* **21**, 591–601 (2007).
6. Talele, T. T., Khedkar, S. A. & Rigby, A. C. Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr. Top. Med. Chem.* **10**, 127–141 (2010).
7. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
8. Sabe, V. T. et al. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur. J. Med. Chem.* **224**, 113705 (2021).
9. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* **38**, 1686–1701 (2015).
10. Jayatunga, M. K., Xie, W., Ruder, L., Schulze, U. & Meier, C. AI in small-molecule drug discovery: A coming wave. *Nat. Rev. Drug Discov.* **21**, 175–176 (2022).
11. Xiang, M., Cao, Y., Fan, W., Chen, L. & Mo, Y. Computer-aided drug design: lead discovery and optimization. *Comb. Chem. high. throughput Screen.* **15**, 328–337 (2012).
12. Pârvu, L. QSAR-a piece of drug design. *J. Cell. Mol. Med.* **7**, 333–335 (2003).
13. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
14. Liu, W. et al. Serial femtosecond crystallography of G protein–coupled receptors. *Science* **342**, 1521–1524 (2013).
15. Jones, N. Crystallography: Atomic secrets. *Nature* **505**, 602–603 (2014).
16. Fernandez-Leiro, R. & Scheres, S. H. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537**, 339–346 (2016).
17. Renaud, J.-P. et al. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nat. Rev. Drug Discov.* **17**, 471–492 (2018).
18. Grygorenko, O. O. et al. Generating multibillion chemical space of readily accessible screening compounds. *Iscience* **23**, 101681–101681 (2020).
19. Hansson, T., Oostenbrink, C. & van Gunsteren, W. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **12**, 190–196 (2002).
20. Baig, M. H. et al. Insight into the effect of inhibitor resistant S130G mutant on physico-chemical properties of SHV type beta-lactamase: A molecular dynamics study. *PLoS One* **9**, e112456 (2014).
21. Freddolino, P. L., Harrison, C. B., Liu, Y. & Schulten, K. Challenges in protein-folding simulations. *Nat. Phys.* **6**, 751–758 (2010).
22. Voter, A. F. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **78**, 3908 (1997).
23. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919–11929 (2004).
24. Hamelberg, D., de Oliveira, C. A. F. & McCammon, J. A. Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. *J. Chem. Phys.* **127**, 155102–155102 (2007).
25. Acharya, K. R., Sturrock, E. D., Riordan, J. F. & Ehlers, M. R. Ace revisited: a new target for structure-based drug design. *Nat. Rev. Drug Discov.* **2**, 891–902 (2003).
26. Lipscomb, W. N. Structure and mechanism in the enzymic activity of carboxypeptidase A and relations to chemical sequence. *Acc. Chem. Res.* **3**, 81–89 (1970).
27. Li, Z. L. & Buck, M. Beyond history and "on a roll": The list of the most well-studied human protein structures and overall trends in the protein data bank. *Protein Sci.* **30**, 745–760 (2021).
28. Nannenga, B. L. & Gonen, T. The cryo-EM method microcrystal electron diffraction (MicroED). *Nat. Methods* **16**, 369–379 (2019).
29. Congreve, M., de Graaf, C., Swain, N. A. & Tate, C. G. Impact of GPCR structures on drug discovery. *Cell* **181**, 81–91 (2020).
30. Greenfield, L. J. Jr Molecular mechanisms of antiseizure drug activity at GABAA receptors. *Seizure* **22**, 589–600 (2013).
31. Rogawski, M. A., Löscher, W. & Rho, J. M. Mechanisms of action of antiseizure drugs and the ketogenic diet. *Cold Spring Harb. Perspect. Med.* **6**, a022780 (2016).
32. Santos, R. et al. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2017).
33. 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).
34. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).
35. Ballante, F., Kooistra, A. J., Kampen, S., de Graaf, C. & Carlsson, J. Structure-based virtual screening for ligands of G protein–coupled receptors: what can molecular docking do for you? *Pharmacol. Rev.* **73**, 1698–1736 (2021).
36. Bender, B. J. et al. A practical guide to large-scale docking. *Nat. Protoc.* **16**, 4799–4832 (2021).
37. Lyu, J., Irwin, J. J. & Shoichet, B. K. Modeling the expansion of virtual screening libraries. *Nat. Chem. Biol.* **19**, 712–718 (2023).
38. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
39. Stein, R. M. et al. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **579**, 609–614 (2020).

40. Alon, A. et al. Structures of the σ2 receptor enable docking for bioactive ligand discovery. *Nature* **600**, 759–764 (2021).
41. Gorgulla, C. et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).
42. Sadybekov, A. A. et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022).
43. Tomberg, A. & Boström, J. Can easy chemistry produce complex, diverse, and novel molecules? *Drug Discov. Today* **25**, 2174–2181 (2020).
44. Gorgulla, C. et al. A multi-pronged approach targeting SARS-CoV-2 proteins using ultra-large virtual screening. *Iscience* **24**, 102021 (2021).
45. Patel, H. et al. SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci. data* **7**, 384 (2020).
46. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *nature* **267**, 585–590 (1977).
47. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.* **9**, 1–9 (2011).
48. Lin, J.-H., Perryman, A. L., Schames, J. R. & McCammon, J. A. Computational drug design accommodating receptor flexibility: the relaxed complex scheme. *J. Am. Chem. Soc.* **124**, 5632–5633 (2002).
49. Amaro, R. E., Baron, R. & McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. -Aided Mol. Des.* **22**, 693–705 (2008).
50. Lins, R. D. et al. Molecular dynamics studies on the HIV-1 integrase catalytic domain. *Biophys. J.* **76**, 2999–3011 (1999).
51. Goldgur, Y. et al. Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proc. Natl Acad. Sci.* **95**, 9150–9154 (1998).
52. Goldgur, Y. et al. Structure of the HIV-1 integrase catalytic domain complexed with an inhibitor: a platform for antiviral drug design. *Proc. Natl Acad. Sci.* **96**, 13040–13043 (1999).
53. Schames, J. R. et al. Discovery of a novel binding trench in HIV integrase. *J. Med. Chem.* **47**, 1879–1881 (2004).
54. Hazuda, D. J. et al. A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. *Proc Natl Acad Sci.* **101**, 11233–11238 (2004).
55. Summa, V. et al. Discovery of raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection. *J. Med. Chem.* **51**, 5843–5855 (2008).
56. Voter, A. F., Montalenti, F. & Germann, T. C. Extending the time scale in atomistic simulation of materials. *Annu. Rev. Mater. Res.* **32**, 321–346 (2002).
57. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **1**, 826–843 (2011).
58. van de Walle, A. Simulations provide a rare look at real melting. *Science* **346**, 704–705 (2014).
59. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).
60. Huber, G. A. & Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys. J.* **70**, 97–110 (1996).
61. Voter, A. F. Parallel replica method for dynamics of infrequent events. *Phys. Rev. B* **57**, R13985 (1998).
62. Voter, A. F. in *Radiation effects in solids* 1-23 (Springer, 2007).
63. Miao, Y., Feher, V. A. & McCammon, J. A. Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *J. Chem. Theory Comput.* **11**, 3584–3595 (2015).
64. Pang, Y. T., Miao, Y., Wang, Y. & McCammon, J. A. Gaussian accelerated molecular dynamics in NAMD. *J. Chem. Theory Comput.* **13**, 9–19 (2017).
65. Miao, Y. et al. Accelerated structure-based design of chemically diverse allosteric modulators of a muscarinic G protein-coupled receptor. *Proc. Natl Acad. Sci.* **113**, E5675–E5684 (2016).
66. Seitz, C. et al. Targeting tuberculosis: Novel scaffolds for inhibiting cytochrome bd oxidase. *J. Chem. Inf. Model.* (2024).
67. Wong, C. F. & McCammon, J. A. J. Computer simulation and the design of new biological molecules. *Isr. J. Chem.* **27**, 211–215 (1986).
68. Wong, C. F. & McCammon, J. A. Dynamics and design of enzymes and inhibitors. *J. Am. Chem. Soc.* **108**, 3830–3832 (1986).
69. Jorgensen, W. L., Buckner, J. K., Boudon, S. & Tirado-Rives, J. Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *J. Chem. Phys.* **89**, 3742–3746 (1988).
70. Gilson, M. K., Given, J. A., Bush, B. L. & McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **72**, 1047–1069 (1997).
71. Ghanakota, P. et al. Combining cloud-based free-energy calculations, synthetically aware enumerations, and goal-directed generative machine learning for rapid large-scale chemical exploration and optimization. *J. Chem. Inf. Model.* **60**, 4311–4325 (2020).
72. Konze, K. D. et al. Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *J. Chem. Inf. Model.* **59**, 3782–3793 (2019).
73. Roos, K. et al. OPLS3e: Extending force field coverage for drug-like small molecules. *J. Chem. Theory Comput.* **15**, 1863–1874 (2019).
74. Abel, R., Wang, L., Mobley, D. L. & Friesner, R. A. A critical review of validation, blind testing, and real-world use of alchemical protein-ligand binding free energy calculations. *Curr. Top. Med. Chem.* **17**, 2577–2585 (2017).
75. Abel, R., Wang, L., Harder, E. D., Berne, B. & Friesner, R. A. Advancing drug discovery through enhanced free energy calculations. *Acc. Chem. Res.* **50**, 1625–1632 (2017).
76. Kuhn, B. et al. Prospective evaluation of free energy calculations for the prioritization of cathepsin L inhibitors. *J. Med. Chem.* **60**, 2485–2497 (2017).
77. Abel, R. & Bhat, S. In *Annual Reports in Medicinal Chemistry* Vol. 50 237-262 (Elsevier, 2017).
78. 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).
79. Kuhn, M. et al. Assessment of binding affinity via alchemical free-energy calculations. *J. Chem. Inf. Model.* **60**, 3120–3130 (2020).
80. Gapsys, V. et al. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **11**, 1140–1152 (2020).
81. Song, L. F., Lee, T.-S., Zhu, C., York, D. M. & Merz, K. M. Jr Using AMBER18 for relative free energy calculations. *J. Chem. Inf. Model.* **59**, 3128–3135 (2019).
82. Wang, L., Chambers, J. & Abel, R. Protein–ligand binding free energy calculations with FEP+. *Biomolecular simulations: methods and protocols*, 201-232 (2019).
83. Schindler, C. E. et al. Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.* **60**, 5457–5474 (2020).
84. Mortier, J. et al. Computationally empowered workflow identifies novel covalent allosteric binders for KRASG12C. *ChemMedChem* **15**, 827–832 (2020).
85. Majellaro, M. et al. 3, 4-Dihydropyrimidin-2 (1 H)-ones as antagonists of the human A2B Adenosine receptor: optimization, structure–activity relationship studies, and enantiospecific recognition. *J. Med. Chem.* **64**, 458–480 (2020).

86. Freedman, H. et al. Application of molecular dynamics simulations to the design of nucleotide inhibitors binding to norovirus polymerase. *J. Chem. Inf. Model.* **60**, 6566–6578 (2020).
87. Takahashi, Y. et al. Stomatal CO2/bicarbonate sensor consists of two interacting protein kinases, Raf-like HT1 and non-kinase-activity requiring MPK12/MPK4. *Sci. Adv.* **8**, eabq6161 (2022).
88. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
89. Kreitz, J. et al. Programmable protein delivery with a bacterial contractile injection system. *Nature* **616**, 357–364 (2023).
90. Lim, Y. et al. In silico protein interaction screening uncovers DONSON's role in replication initiation. *Science* **381**, eadi3448 (2023).
91. Mosalaganti, S. et al. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **376**, eabm9506 (2022).
92. Anand, N. & Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019* (2022).
93. Yang, Z., Zeng, X., Zhao, Y. & Chen, R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct. Target. Ther.* **8**, 115 (2023).
94. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
95. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
96. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, 2022.2007. 2021.500999. https://doi.org/10.1101/2022.07.21.500999 (2022).
97. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. computational Chem.* **31**, 455–461 (2010).
98. Miller, E. B. et al. Reliable and accurate solution to the induced fit docking problem for protein–ligand binding. *J. Chem. Theory Comput.* **17**, 2630–2639 (2021).
99. Lyu, J. et al. AlphaFold2 structures guide prospective ligand discovery. *Science*, eadn6354 (2024).
100. Liu, Z. et al. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).
101. Gaudelet, T. et al. Utilizing graph machine learning within drug discovery and development. *Brief. Bioinforma.* **22**, bbab159 (2021).
102. Son, J. & Kim, D. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. *PloS one* **16**, e0249404 (2021).
103. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Improving detection of protein-ligand binding sites with 3D segmentation. *Sci. Rep.* **10**, 5035 (2020).
104. Jiménez, J., Skalic, M., Martinez-Rosell, G. & De Fabritiis, G. K deep: protein-ligand absolute binding affinity prediction via 3d-convolutional neural networks. *J. Chem. Inf. Model.* **58**, 287–296 (2018).
105. Jones, D. et al. Improved protein-ligand binding affinity prediction with structure-based deep fusion inference. *J. Chem. Inf. Model.* **61**, 1583–1592 (2021).
106. Volkov, M. et al. On the frustration to predict binding affinities from protein-ligand structures with deep neural networks. *J. Med. Chem.* **65**, 7946–7958 (2022).
107. MolSoft. (http://molsoft.com/RIDE.html, 2022).
108. Durrant, J. D. & McCammon, J. A. NNScore: a neural-network-based scoring function for the characterization of protein− ligand complexes. *J. Chem. Inf. Model.* **50**, 1865–1871 (2010).
109. 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).
110. Chen, W. et al. Enhancing hit discovery in virtual screening through absolute protein-ligand binding free-energy calculations. *J. Chem. Inf. Model.* **63**, 3171–3185 (2023).
111. Khalak, Y. et al. Alchemical absolute protein-ligand binding free energies for drug design. *Chem. Sci.* **12**, 13958–13971 (2021).
112. Cournia, Z. et al. Rigorous free energy simulations in virtual screening. *J. Chem. Inf. Model.* **60**, 4153–4169 (2020).
113. Bender, A. & Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discov. Today* **26**, 511–524 (2021).
114. Davies, M. et al. Improving the accuracy of predicted human pharmacokinetics: lessons learned from the AstraZeneca drug pipeline over two decades. *Trends Pharmacol. Sci.* **41**, 390–408 (2020).
115. Schneckener, S. et al. Prediction of oral bioavailability in rats: Transferring insights from in vitro correlations to (deep) machine learning models using in silico model outputs and chemical structure parameters. *J. Chem. Inf. Model.* **59**, 4893–4905 (2019).
116. Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
117. Keiser, M. J. et al. Predicting new molecular targets for known drugs. *Nature* **462**, 175–181 (2009).
118. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331 (2016).
119. Cichońska, A. et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. *Nat. Commun.* **12**, 3307 (2021).


## Acknowledgements

This work has been supported by funding provided by UCSD. The authors' research computing has been supported by UCSD's SDSC TSCC systems.

## Author contributions

H.W. and J.A.M. both wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to Haixin Wei or J. Andrew McCammon.