Structure-Based Drug Discovery with Deep Learning**

ChemBioChem

Review
doi.org/10.1002/cbic.202200776

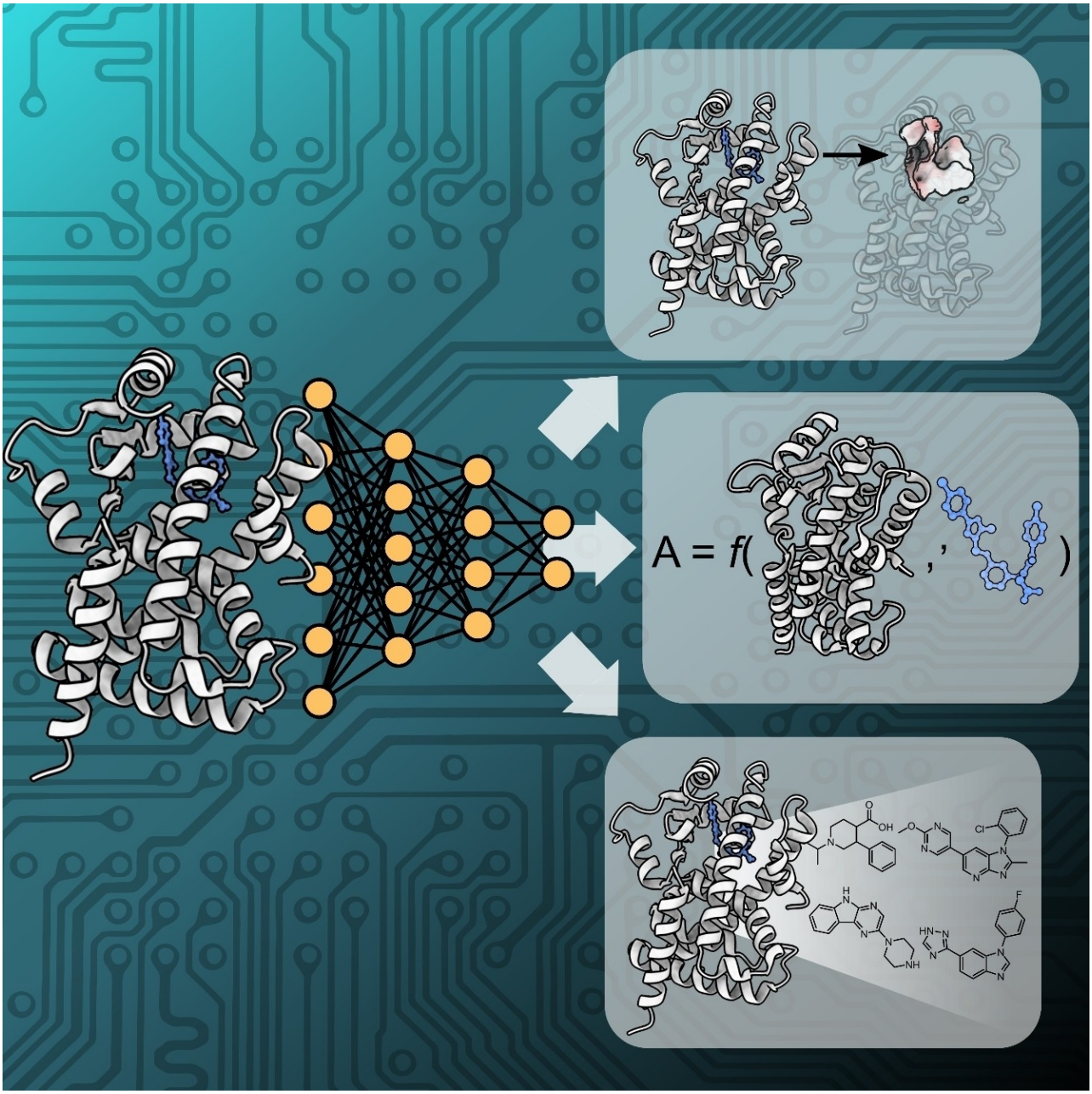
Chemistry Europe
European Chemical Societies Publishing

www.chembiochem.org

VIP *Very Important Paper*

Structure-Based Drug Discovery with Deep Learning**

R. Özçelik[+],[a, b] D. van Tilborg[+],[a, b] J. Jiménez-Luna,[c] and F. Grisoni*[a, b]



Artificial intelligence (AI) in the form of deep learning has promise for drug discovery and chemical biology, for example, to predict protein structure and molecular bioactivity, plan organic synthesis, and design molecules *de novo*. While most of the deep learning efforts in drug discovery have focused on ligand-based approaches, structure-based drug discovery has the potential to tackle unsolved challenges, such as affinity prediction for unexplored protein targets, binding-mechanism elucidation, and the rationalization of related chemical kinetic properties. Advances in deep-learning methodologies and the availability of accurate predictions for protein tertiary structure advocate for a *renaissance* in structure-based approaches for drug discovery guided by AI. This review summarizes the most prominent algorithmic concepts in structure-based deep learning for drug discovery, and forecasts opportunities, applications, and challenges ahead.

1. Introduction

Deep learning – a subfield of artificial intelligence (AI) based on multilayer neural networks[1] – has gained remarkable traction in science and technology, for example, to advance mathematics,[2,3] investigate galaxies,[4] and generate realistic images.[5] Chemistry and biology have witnessed several AI breakthroughs, for instance, in protein structure prediction,[6,7] chemical synthesis planning,[8,9] and atomistic simulations,[10,11] Drug discovery has particularly benefited from the advent of deep learning,[12,13] achieving success in molecule prioritization and automated *de novo* design,[14–17] Here, deep learning can accelerate the navigation of the extremely vast chemical space of drug-like molecules[18] in search for potential therapeutics, and complement resource- and time-intensive high-throughput screening campaigns.[19] Most deep learning studies have focused on ligand-based approaches,[12] which leverage solely the structural information of small-molecule ligands to provide predictions. For these applications, numerous systematic studies[20,21] and experimental proofs-of-concept[16,17,22] have been published. On the other hand, structure-based deep-learning approaches – which leverage information on the target protein – have not found parallel interest yet.

Structure-based drug discovery (SBDD) methods augmented with AI are arguably a more complex and a higher-potential endeavor compared to their ligand-based counterparts. Numerous marketed drugs have been identified by "traditional" SBDD (*e.g.*, HIV-1 protease inhibitors.[23] the thymidylate synthase inhibitor raltitrexed,[24] and the antibiotic norfloxacin[25]). Accelerating SBDD with deep learning can help address existing drug discovery challenges, such as polypharmacology by design,[26] selectivity optimization,[27] activity cliff prediction,[28] and target deorphanization.[29] Deep learning does not require explicit feature engineering and can thus be applied to learn directly from molecular representations of both ligands and proteins. This is particularly relevant for SBDD, where engineering numerical features for complex molecular entities like proteins[30] is inevitably more laborious than for small molecules.[31] Therefore, deep learning for SBDD bears an untapped potential to capture highly nonlinear structure-activity relationships and has recently started to show its promise. Accurate protein structure prediction efforts like AlphaFold[6,7] are expected to further accelerate computer-assisted SBDD. Deep learning for SBDD is still in its infancy but is moving forward at a fast pace, and its relevance in the years to come is expected to increase.

This review provides a comprehensive overview of how deep learning can be leveraged for SBDD, and how to incorporate protein information at different levels of complexity (*e.g.*, amino-acid sequence, and/or tertiary structure). After addressing how proteins can be represented for deep learning, we address current state-of-the-art methods for structure-based drug discovery, with a particular focus on drug-target interaction prediction, binding site detection, and *de novo* design (Figure 1). Finally, we discuss current limitations and research gaps, along with foreseen future directions and opportunities. A glossary of selected terms can be found in Table 1.

2. Representing Proteins for Deep Learning

The design of deep-learning approaches for SBDD is inherently more intricate than for ligand-based approaches, due to the need to represent protein information at different levels of complexity. Proteins are large polypeptide chains that are hierarchically organized into:[35] a) *primary structure*, referring to the sequential arrangement of amino acids along the polypeptide chain, b) *secondary structure*, capturing the occurrence of alpha-helices and beta-pleated sheets along the protein sequence, and c) *tertiary structure*, capturing how proteins fold in the three-dimensional space. Such complexity is reflected in the various protein representations used for deep learning (Figure 2):

[a] *R. Özçelik,[+] D. van Tilborg,[+] F. Grisoni*
Institute for Complex Molecular Systems and Dept. Biomedical Engineering, Eindhoven University of Technology
5612 AZEindhoven (The Netherlands)
E-mail: f.grisoni@tue.nl
[b] *R. Özçelik,[+] D. van Tilborg,[+] F. Grisoni*
Alliance TU/e, WUR, UU, UMC,
Centre for Living Technologies
3584 CBUtrecht (The Netherlands)
[c] *J. Jiménez-Luna*
AI4Science,
Microsoft Research
Cambridge, CB1 2FB (UK)
[+] *These authors contributed equally.*
[**] *A previous version of this manuscript has been deposited on a preprint server (https://arxiv.org/abs/2212.13295).*
This article is part of the Special Collection ChemBioTalents2022. Please see our homepage for more articles in the collection.

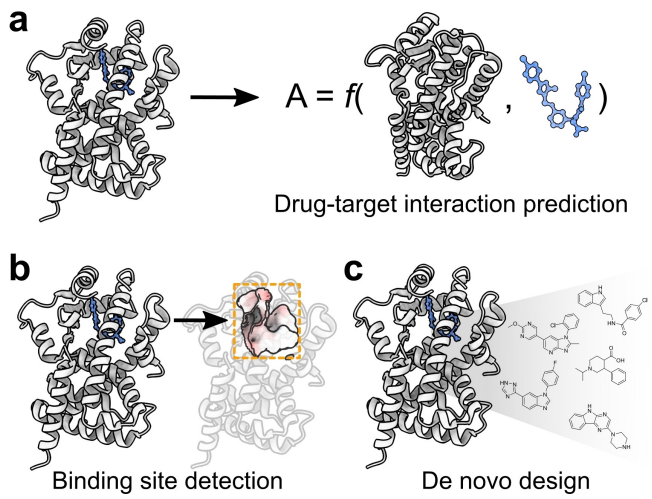
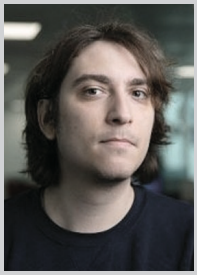


Figure 1. Structure-based drug discovery tasks discussed in this review: a) *drug-target interaction prediction*, which aims to predict the affinity between a protein and a ligand by using the structural information of both molecular entities; b) *binding site detection*, which aims to identify druggable cavities in the protein structure, c) *de-novo design*, aiming to design bioactive molecules from scratch by using the information of a protein target.

- *Primary (amino-acid) sequence.* The amino acid sequence is specified starting from the amino-terminal end (N-terminus) and ending at the carboxyl-terminal (C-terminus) end. For deep learning purposes, the primary sequence is often represented as a character string, where each letter represents one of the 20 naturally occurring amino acids (*e.g.*, "A-I-R" corresponds to alanine, isoleucine, and arginine). These representations are at the core of established protein featurization techniques, such as ProtVec,[36] Evolutionary scale modeling (ESM),[36] unified protein representations (UniRep),[37] ProteinBERT,[38] SeqVec,[39] and ProtTrans.[40] Although less frequently encountered,[41,42] the primary sequence can also be represented as a graph whose nodes are amino acids (featurized by type or corresponding physicochemical features), and whose edges capture their adjacency in the chain.
- *Tertiary (3D) structure.* The three-dimensional shape of a protein (tertiary structure) is determined by the interactions among its side chains, and features a certain degree of conformational plasticity.[43] The protein structure contains key information for SBDD, as it relates to protein function,[44,45] and it determines ligand binding.[46] Moreover, inducing conformational changes is often the goal of drug discovery.[43] Several ways exist to learn from tertiary structures with deep learning. Early approaches[47,48] have used grid-based voxel representations (Table 1) to capture the spatial distribution of the protein's physicochemical or pharmacophore properties. While these representations are suited for well-established deep learning architectures (*e.g.*, convolutional neural networks, CNNs), many voxels representing empty space do not carry relevant information and increase computational costs at higher spatial grid resolutions. Other approaches[49] represent proteins as molecular graphs in combination with graph neural networks,[50] where

Rıza Özçelik is a Ph.D. candidate in the Molecular Machine Learning team at the Eindhoven University of Technology under the supervision of Dr. Francesca Grisoni. He received his MSc. degree from the Department of Computer Engineering at Boğaziçi University, Turkey, where he studied structure-based drug-target affinity prediction. His current research focuses on developing novel generative deep-learning approaches for de novo chemical design.

Derek van Tilborg is a Ph.D. candidate in the Molecular Machine Learning team led by Dr. Francesca Grisoni at the Eindhoven University of Technology. Derek holds a MSc degree in bioinformatics from Wageningen University & Research and has a background in biomedical research. He currently works on developing machine learning methods for molecular property prediction, aiming to bridge the gap between predictions and experiments in drug discovery. His research interests are focused on graph neural networks and active learning.



José Jiménez-Luna is a Senior Researcher at Microsoft Research Cambridge interested in applications at the interface between cheminformatics and statistical learning. Prior to joining MSR, he was a postdoctoral fellow in Gisbert Schneider's laboratory at ETH Zurich, funded by Boehringer Ingelheim Pharma, where he became interested in explainable artificial intelligence methods in the context of drug design and chemical property prediction. He obtained his Ph.D. in 2019 at Gianni de Fabritiis' biophysics laboratory at Pompeu Fabra University, on pioneering applications of convolutional neural networks to 3D atomistic systems..

Francesca Grisoni is an Assistant Professor at the Eindhoven University of Technology. After receiving her Ph.D. in 2016 at the University of Milano-Bicocca (Prof. R. Todeschini) on machine learning for toxicology, she worked as a data scientist and biostatistical consultant for the pharmaceutical industry. In 2018, she joined the group of Gisbert Schneider (ETH Zurich) as a postdoctoral fellow, working on deep learning for de novo design. Francesca's team focuses on developing novel deep learning methods for drug discovery, at the interface between computation and wet-lab experiments.

Table 1. Glossary of selected terms, reporting key definitions from chemical biology and machine learning.

Term	Description
Binding site	Protein region that is responsible for the interaction with another molecule (e.g., small-molecule inhibitors, activators, and other proteins).
Generative deep learning	Deep learning methods that aim to model the underlying data distribution of a given set of samples and, by sampling from the modeled distribution, generate new data points without the need for explicit hard-coded design rules.[32]
Geometric deep learning	Umbrella term to identify neural network architectures that incorporate and process symmetry information in their design.[33]
Featurization	Conversion of various types of data into numerical data (features) for machine learning.
Ligand	Any molecule that binds to a protein with high affinity.
Molecular descriptors	Numerical features obtained from a molecular representation with the goal of capturing pre-defined chemical information.[31]
Molecular docking	Computational procedure used to predict the predominant three-dimensional binding mode(s) of a molecule w.r.t. another (macro)molecule it binds to. These typically involve the use of a conformational pose search method and a scoring function.[34]
Protein	(Macro)molecule consisting of amino acid residues joined by peptide bonds.
Reinforcement learning	Subfield of machine learning whose goal is to study the behavior of agents that learn a sequence of actions that maximize a cumulative reward within a specific environment.
Simplified molecular input line entry system (SMILES)	String-based chemical notation capturing two-dimensional molecular information, in which letters are used to represent atoms, whereas symbols and numbers encode bond types, connectivity, branching, and stereo-chemistry.
Transfer learning	A machine learning method where a model trained on one task is reused as the starting point for a model on a second, related, task.
Voxel	A volumetric pixel.

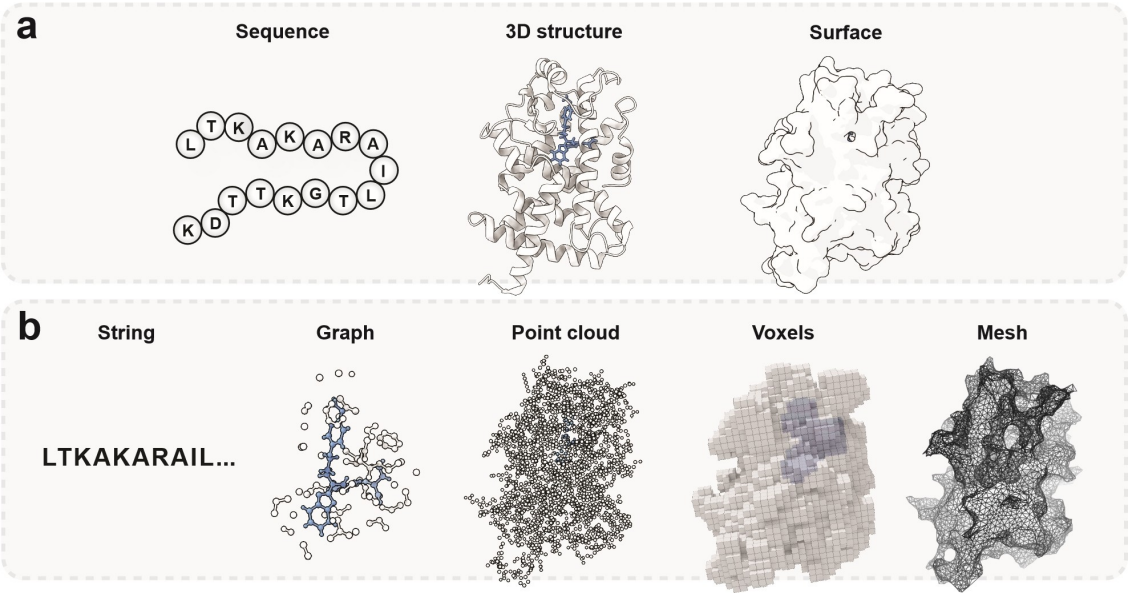


Figure 2. Representing proteins for deep learning. a) Structural hierarchy of protein information: *(primary) amino-acid sequence*, referring to the sequential arrangement of amino acids along the polypeptide chain; *(tertiary) 3D structure*, capturing protein folding in the three-dimensional space; *protein surface*, delimiting solvent-accessible and inaccessible regions. Each level of information is characterized by a different availability of data.[2] b) Protein representations for deep learning, capturing information on the protein sequence (*strings*), the 3D structure (*graphs*, *voxels* and *point clouds*), and the surface (*point clouds* and *meshes*). Each representation is suited to different neural network architectures (Figure 3).

each atom is a node and each bond is an edge. Depending on the chosen level of coarse-graining, often only backbone atoms are chosen to correspond to nodes, while edges often represent geometrical proximity in the coordinate space rather than direct chemical bonds.[51] Edges and bonds can be characterized by additional geometrical and/or physico-chemical properties.

- *Protein surface*. The protein surface is usually defined as the separation between solvent-accessible and inaccessible regions[52,53] (Figure 2a) and it plays a key role in the protein interactions with (macro)molecular entities. Protein surfaces are usually represented as either meshes (*i.e.*, a set of polygons capturing the location of the surface, whose vertices can be described by a 2D grid or a 3D graph structure) or point clouds (*i.e.*, graphs whose nodes describe the location of the surface at a certain resolution). Although often computed from the 3D structure, the surface representation might better reflect the physicochemical features

responsible for the interaction with other (macro)molecules, as well as aspects of protein function that go beyond sequence similarity.

Small-molecule ligands can be represented in analogous ways to protein structures. The most used representations are: a) molecular strings (*e.g.*, SMILES strings[54]), which capture 2D information (atom occurrence and connectivity), b) 2D and 3D molecular graphs (based on the availability of experimentally determined or computed conformational information), and c) molecular surfaces. An in-depth description of small-molecule representations and corresponding deep-learning approaches can be found in a recent work.[55]

3. Data for Structure-Based Drug Discovery

Deep learning models are notoriously "data hungry". In this context, not only does the chosen protein representation affect the type and quality of chemical information captured, but it also determines the number of data points available for training (Table 2). Primary sequence data are abundant (*e.g.*, more than 60 M sequences are available on Uniprot[56]) but lack information about the spatial configuration of atoms, which determines the binding pose of ligands. On the other hand, 3D protein structures contain richer information for many drug discovery purposes, but are relatively scarce. Typically obtained through expensive experimental methods like X-ray diffraction or NMR spectroscopy,[57] they are available in the order of hundreds of thousands (Table 2). Luckily, the cost of protein structure determination has been steadily decreasing over time, thanks to the advent of newer experimental techniques like cryogenic electron microscopy.[58] Despite these advances, 3D protein structures still come with their own caveats. Obtaining high-quality protein structures is oftentimes resource-intensive and challenging for several targets,[59] such as disordered and membrane proteins. Furthermore, it is currently difficult to co-crystallize weakly binding ligands, which results in highly unbalanced data for model training.[60,61] Deep learning breakthroughs in protein structure prediction[6,62] bear promise to bridge the gap in data availability, by making thousands of predicted protein structures available for downstream tasks.[63–66] Despite this, the quality of machine-learning-based structure predictions is known to depend on several factors, such as the protein length and its flexibility, as well as the presence of similar structures in the training set.[67,68]

(Macro)molecules are dynamical entities and are always interconverting between a variety of conformations with varying energies,[69] with key implications in drug-target interaction.[70] Considering protein (and ligand) conformational flexibility is thus key to understanding several biological processes.[71] However, information on dynamics is currently missing from experimentally determined datasets (Table 2) as well as from structures predicted by AI.[72] In this context, molecular dynamics simulations can provide insights into (un)binding and conformational changes at a spatial and temporal resolution that is not available experimentally or from structure prediction. However, to date, these models might have prohibitive computational costs.

Finally, ligand potency databases currently suffer from notable biases,[73] as existing literature tends to over-report analogues and binding compounds.[74] Moreover, public repositories inherently suffer from assay heterogeneity and experimental noise. Researchers using these resources should be aware of these limitations and curate their data with care, even

Table 2. Summary of selected datasets for structure-based deep learning. Dataset name, description and number of entries (updated as of June 2022) are provided.

Dataset	Description	No. entries	Link (if available)
Protein Data Bank (PDB)[76]	Structural data of biological macromolecules.	189,735 structures	rcsb.org
scPDB[77]	Druggable binding sites and ligands extracted from the PDB.	4782 protein structures and 6326 ligands.	bioinfo-pharma.u-strasbg.fr/scPDB
BioLip[78]	Semi-manually curated ligand-protein interactions.	573,225 entries.	zhanggroup.org/BioLiP
PDBbind[56]	Protein-ligand co-complexes and associated affinities extracted from the PDB.	23,496 complexes.	pdbbind.org.cn
UniProt[56]	Protein sequence and functional information.	>60 million sequences.	uniprot.org
AlphaFold Protein Structure Database[63]	Predictions of protein structures by AlphaFold v 2.0.[6]	992,316 predicted structures.	alphafold.ebi.ac.uk
AlphaFill[64]	Common ligands and cofactor transplants for AlphaFold models.	12,029,789 transplants.	alphafill.eu
Binding MOAD[79,80]	X-ray crystal structures with bound ligands and experimental binding affinities.	41,409 protein-ligand complexes, and 15,223 binding measurements.	bindingmoad.org
Directory of Useful Decoys (DUD–E)[81]	Directory of decoys designed to benchmark molecular docking programs.	22,886 active molecules and affinities on 102 targets; 59 decoys per compound.	dude.docking.org
BindingDB[82] validation sets	Binding affinities of protein-ligand pairs curated from the literature.	~1200 series with at least 1 cocrystal available in each.	bindingdb.org
BigBind[83]	Associated protein structures to ChEMBL[84] assay data via Pocketome.[85]	818,995 activities with associated protein structures.	Brocidiacono *et al.*, 2022
KIBA[86]	Bioactivity measurements of compounds against kinases.	246,088 measurements.	Tang *et al.*, 2014
Davis[87]	Binding affinities (K_d values) of inhibitors against kinases.	30,056 measurements.	Davis *et al.*, 2011

though it has been noted that fully accounting for aspects such as experimental noise and heterogeneity is practically impossible.[75]

4. Deep Learning for Structure-Based Drug Discovery

This section aims to provide a concise overview of SBDD approaches fueled by deep learning. SBDD will be considered in its broader sense, that is, not only limited to 3D protein structure but also including sequence and surface representations. Each of these representations is suited to different neural network architectures. Although many "flavors" of deep learning for molecular representations exist,[55] some of the most well-established architectures for SBDD are:

- *Recurrent neural networks* (RNNs; *e.g.*, with long-short term memory cells[88]) are a commonly chosen architecture to process the primary sequence of a protein. RNNs incorporate feedback connections that allow information in the previous inputs to flow into the subsequent inputs. The feedback mechanism behaves as a "learned memory" in the architecture and enables capturing sequential structure of the input (Figure 3a).
- *Convolutional neural networks* (CNNs) are a powerful architecture when paired with voxelized representations to capture spatial dependencies. CNNs apply learnable filters to the input and excel in capturing local patterns, which renders them suited to binding affinity prediction and pocket detection (Figure 3b).
- *Graph neural networks* (GNNs) operate on molecular graphs (*e.g.*, atoms and their interactions) and can capture the structural and functional relationships between, as well as within, atoms belonging to one or more molecular complexes (Figure 3c). The representation flexibility of GNNs makes them applicable to a wide variety of tasks in structure-based drug discovery.

Deep-learning approaches can be applied to different tasks, based on the envisioned application and the utilized input representation. In what follows, we focus on three key tasks (Figure 1), namely binding site detection, drug-target interaction prediction, and structure-based *de novo* design. For each task, selected deep-learning approaches are described through the lenses of the protein representation they rely on (sequence, structure, or surface). A summary of selected deep learning studies is reported in Table 3.

4.1. Drug-target interaction prediction

The identification of interactions between molecules and macromolecular targets is a key step in drug discovery, drug repurposing, and off-target activity prediction. Drug-target interaction (DTI) prediction aims to predict the bioactivity (*e.g.*, binding affinity) of a given set of molecules on one or more macromolecular targets, by leveraging both protein and ligand information. Given the complexity of ligand-protein interactions and of engineering suitable molecular features for DTI, it is no surprise that this topic has found a widespread application of deep learning techniques.[132] In what follows, deep learning models developed for DTI prediction are categorized on the basis of the protein representation they rely on.

- *Sequence-based approaches.* Sequence-based DTI prediction models use amino-acid sequences in combination with additional ligand representations to provide predictions. One of the earliest approaches, DeepDTA,[89] applied 1D CNNs to simultaneously learn from string representations of both

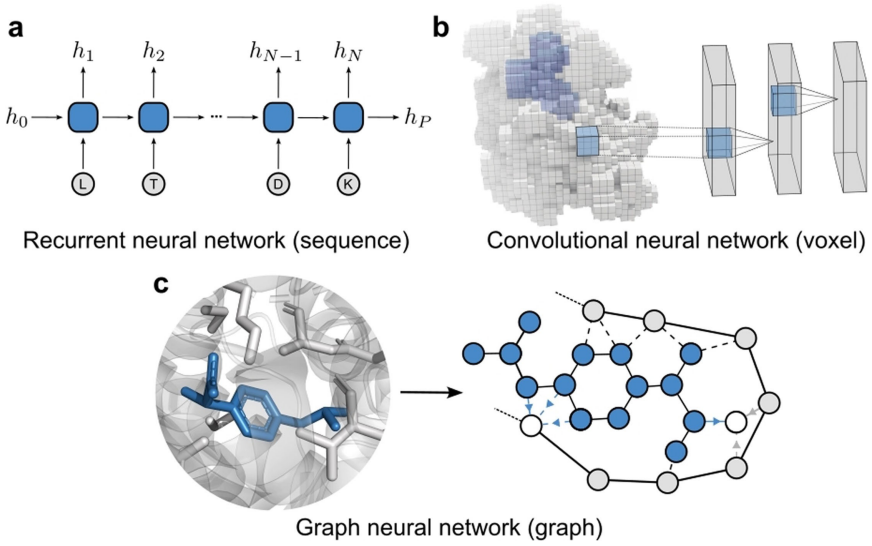


Figure 3. Selected neural network architectures commonly employed for SBDD. a) Recurrent neural networks can be used to learn from the primary sequence of a protein, as well as other sequential molecular representations. b) Convolutional neural networks (CNN) capture 3D spatial information and are commonly used to learn from voxelized protein representations. c) Graph neural networks (GNNs) operate on molecular graphs and are often used to learn from interatomic interactions within and between molecules.

Table 3. Selected deep-learning approaches applied to SBDD. Models are categorized by task and frequently adopted molecular representations.

Task	Description	Protein representation
Drug-target interaction prediction	Predict the interaction between one or more proteins and one or more ligands.	Amino-acid sequence[89–94] 3D structure[47,95–98]
Docking	Determination of a ligand pose within a target binding site.	3D structure[99–101]
Binding site detection	Identification and/or localization of functional protein binding sites.	Amino-acid sequence[90,102–104] 3D structure[48,105–110] Surface[111–113]
De novo design	Generation of ligands with desired properties conditioned on a protein.	Amino-acid sequence[114–116] 3D structure[117–131]

ligands (in the form of SMILES) and protein sequence. This is achieved by first creating separated embeddings and then concatenating them to perform a prediction. Later works have replaced CNNs with other methods such as recurrent neural networks (Figure 3),[90,133] attention-based[90,133,134] or transformer architectures.[91,92,135] Several works have addressed how to improve the representation of protein sequences, for example, by incorporating evolutionary information,[136,137] or protein sequence composition descriptors.[93,138] Ligands are usually represented as strings (*e.g.*, SMILES[89,137] or DeepSMILES[136,139]), fingerprints,[93,140] frequently occurring substructures,[136] or molecular graphs.[138,141,142]

- *3D structure-based approaches.* These models leverage atom coordinates, usually of co-crystallized protein-ligand complexes, for training. Early approaches projected 3D protein-ligand complexes into grids featurized with physiochemical properties, and subsequently applied CNNs for binding affinity prediction.[47,95,143] Later works extended this idea by including more sophisticated features, for example, intermolecular interaction fingerprints[96] and computed molecular energies.[144] 3D grid-based approaches have also been used for lead optimization by predicting relative binding free energies linked to small modifications of ligand structures.[145,146] More recent approaches have replaced grid-based representations with graphs,[49,97,98,147] allowing to explicitly represent atom neighborhoods and connectivity, and apply roto-translational invariant graph neural networks for binding affinity prediction.
- *Surface-based approaches.* Surface-based approaches have found limited application for DTI prediction. OctSurf[148] represents both binding pockets and ligands as surfaces, by partitioning the 3D space recursively into octants and considering only portions containing van der Waals surface points. Non-empty octants, along with their physicochemical and geometric features, are then used as the input to a CNN. Other approaches, such as HoloProt,[149] merge 3D structure (graph) and surface (point cloud) information for task-specific training, for example, enzyme-catalyzed reaction classification and binding affinity prediction.

Another topic of recent interest by the deep learning community is *protein-ligand docking*, which aims to predict the putative binding pose of a ligand upon binding to a macromolecular target (Table 1). Although these methods do not aim to predict the affinity between a ligand-protein pair directly, they can be used as a proxy to elucidate potential mechanisms of interaction. Deep learning has been mostly applied to ligand pose optimization while considering a rigid target structure, although recent approaches have started taking side-chain flexibility into account.[150] Early approaches used protein-ligand interaction fingerprints,[151,152] while successive approaches have leveraged either a voxelized version of the protein structure combined with CNNs[153,154] or graph-based representations with message-passing neural networks[155–158] in lieu of classical scoring functions. Finally, several approaches have attempted to directly predict the ligand binding pose in an end-to-end fashion,[99,100,159] without the need for a classical search algorithm by exploiting advances in equivariant deep learning.

Deep learning has undoubtedly accelerated DTI prediction, thanks to the possibility to represent and learn from protein-ligand complexes more efficiently. However, simpler models based on well-established descriptors might reach comparable performance,[160] due to undesired memorization and hidden bias in ligand-protein interaction data.[60,161–163] Moreover, no relationship has been observed between the complexity of protein and ligand representations and the accuracy of the resulting deep learning models.[160] Thus, more attention should be put on strategies for model evaluation and data selection/splitting procedures to ensure a reliable prediction of DTIs with deep learning.[160,164,165]

4.2. Binding site detection

The identification of druggable binding sites in proteins plays a pivotal role in SBDD, from hit identification and molecule screening to water interaction site prediction[166,167] and mechanism formulation.[168] Over the years, a plethora of methods have been developed for binding site detection,[168–172] for example, via interatomic gap volumes[170] or regions of buried pocket surfaces.[173] Recently, deep learning methods learning directly from "raw" representations of proteins have gained increasing traction to detect binding sites. These approaches can be grouped by the molecular representations they rely on, that is, protein sequence, 3D structure, and surface, as described below:

- *Sequence-based models.* Binding site detection can be performed by predicting which residues of the amino-acid sequence are involved in ligand binding, although sequence-based approaches have found limited application. Early

methods approached binding site detection as a "side-product" of binding affinity prediction, by using explainable AI techniques to highlight relevant residues for a model's prediction.[90,102,103] Few works have addressed binding site detection only.[174,175] Recently, sequence-based binding site detection has been jointly modeled with drug-target affinity prediction, leading to improved performance on both tasks.[104]

- *3D structure-based models*, which use the spatial information of proteins to detect likely binding sites. Early approaches represented the protein structure with voxels featurized with pharmacophore-like properties, along with convolutional neural networks.[48,105] Subsequent works have refined structure-based binding site detection with additional techniques from the computer vision domain, for example, image segmentation.[106,107] BiteNet[176] also used CNN-based approaches but additionally incorporated conformational ensembles of proteins. The approach was later adapted to predict protein-peptide binding sites.[177] Recent approaches have also leveraged AlphaFold and amino-acid level features to predict binding sites,[108,109] as well as graph neural networks to discover cryptic binding pockets.[110]
- *Surface-based models*. Voxelized representations of protein coordinates have several drawbacks,[113,178] and might lead to worse results than working with surfaces alone.[178] Although deeply buried amino acids often affect the properties of the protein surface,[179,180] voxelized methods often carry non-informative voxels that represent empty space and suffer from information coarse-graining due to discretization of the input protein space. For this reason, several methods based on protein surfaces have been developed over the years. These approaches rely on the representation of protein structures as continuous shapes characterized by geometric and physicochemical features to perform a prediction. Geodesic CNNs have been used to determine interaction fingerprints of molecular surfaces, and to predict protein and ligand binding sites.[111] This approach was later expanded to obtain fully learnable protein representations.[112] Alternatively, DeepSurf[113] discretizes the solvent accessible surface using a combination of *k*-means clustering and density reduction.[113]

A recent analysis of computational approaches for protein-ligand binding site recognition[181] has shown DeepSurf[113] to perform remarkably well. Moreover, some non-machine-learning algorithms have been shown to be competitive alternatives to deep learning.[181] One caveat is that all methods compared struggled on shallow binding sites, due to the higher frequency of deep grooves used for model training.[181] Despite the recent progress on binding pocket detection, room for improvement remains, for example, to increase pocket coverage and detect subpockets,[181] and to predict allosteric binding sites.[182]

4.3. Protein-based de novo molecule design

De novo design refers to the generation of novel chemical entities possessing desired properties from scratch[183] and is among the most challenging tasks in computer-assisted drug discovery. Computational algorithms are faced with an incredibly vast "chemical universe", whose cardinality has been estimated between 10^{24} and 10^{100} molecules.[18,184,185] In this context, "brute-force" molecule assembly or enumeration approaches are computationally unfeasible. In recent years, generative deep learning has shown great promise for *de novo* drug design[15,17,186] and to complement traditional approaches based on human-engineered rules.[24,187–189]

Generative deep learning approaches for *de novo* design are usually applied to produce molecules in the form of molecular graphs[190–192] or strings[15–17] (*e.g.*, SMILES). While most *de novo* design approaches are ligand-based,[15,186,193–196] structure-based approaches have recently emerged as a promising research direction,[197] due to their potential to design molecules interacting with pharmacologically relevant targets on demand.

- *Sequence-based approaches.* Sequence-based *de novo* design approaches usually cast the problem into a machine translation task, where high-affinity protein-ligand pairs are considered as sentences in different languages to be matched. To this end, amino-acid sequences and SMILES strings for proteins and ligands are used, respectively. The first-in-kind approach[114] trained a transformer architecture to "translate" amino acid sequences into the SMILES strings of the corresponding ligand. This approach can be used for sequence-conditioned *de novo* design. A recent work used a transformer-based pipeline[115] that combined language models that were pre-trained on large corpora of proteins[198] and small molecules.[199] Another recent sequence-based *de novo* design approach replaced the machine translation formulation with a reinforcement learning setting and conditioned the designs of a ligand-based molecule generator with a drug-target affinity prediction model.[116] The resulting designs showed drug-like properties and promising docking scores on the selected targets.
- *3D structure-based approaches. De novo* design conditioned on the tertiary structure information can usually generate molecules in the form of 3D ligands (molecular graphs) or strings (*e.g.*, SMILES). In the former case, 3D representations of protein-ligand complexes are used as the input to generate novel 3D molecular graphs. As one of the earliest approaches, LigVoxel[117] relied on 3D grids to generate spatial "blobs" of ligand properties such as occupancy, aromaticity, and hydrogen-bond donor/acceptors that match the protein pocket. Later works used diffusion models,[126,127] variational autoencoders,[118] and reinforcement learning[119,120,128] to directly generate ligand conformations inside the binding pocket. Recently, equivariant neural networks coupled with point-cloud representations have been used for molecule optimization, via pocket-based fragment expansion,[200] as well as generative adversarial networks that represent proteins at the atomic level.[129] Compared to 3D graphs, molecular strings are usually easier to generate and might match or outperform graph-based models.[201] A pioneering work of this category leveraged generative adversarial networks[202] to produce SMILES strings conditioned on voxelized protein pockets.[122] A subsequent model

adopted graph neural networks to represent active sites and generated targeted SMILES strings,[124,130] whereas other models enriched 3D information with pharmacophore models to condition the generation for the targeted pocket.[123,131] Recently,[125] a recurrent neural network model has been coupled with ligand-protein interaction fingerprints determined on ligand docking poses for conditioned ligand generation in the form of SMILES strings.

While ligand-based *de novo* design pipelines using deep learning have been experimentally validated in multiple instances,[17,203,204] structure-based *de novo* design has, to date, not been applied prospectively. This aspect represents an important gap for their widespread adoption.

5. Gaps, Opportunities, and Outlook

Deep learning for structure-based drug discovery is gaining increasing traction, as evidenced by the exponential increase in the number of published approaches over the last few years. Breakthroughs in protein structure prediction[7,63] not only exemplify the potential of deep learning in the molecular sciences, but are expected to further propel structure-based drug discovery with AI. Crucially, SBDD promises to tackle drug discovery for new, uncharted protein regions. Such "zero-" and "few-shot" learning frameworks makes AI-driven SBDD a high-risk/high-gain endeavor, expected to advance future drug discovery.

SBDD is arguably more challenging than its ligand-based counterpart, due to the number and structural complexity of chemical entities involved, and aspects like target conformational flexibility.[205] Most of the current SBDD algorithms are agnostic to dynamical information and, in certain instances, might not outperform simpler methods.[160,181] While some elucidation of dynamics has been previously attempted, for example, from AlphaFold-predicted structures[206] or by molecular dynamics,[207] current methods in the realm of SBDD do not routinely use this information. This limitation is exacerbated by the scarcity of open molecular dynamics datasets on protein-ligand complexes. Because dynamics play a crucial role in lead optimization,[208–210] any efforts in this area are likely to substantially accelerate progress in the field.

Geometric deep learning is an emerging research area that is finding one of its major applications in SBDD.[33,211] These approaches attempt to unify neural networks from the perspective of symmetry and topology. Roto-translational invariance/equivariance is particularly relevant for the three-dimensional representations of molecular systems, which is beneficial to limit function search space during training.[55] In this review, we have already touched upon several such approaches.[99,126,127,212] We expect the application of geometric deep learning for 3D structures to further boost AI's ability to model molecular complexes and their interactions,[111,213,214] as well as molecular design[127,215] in the future. On a related note, geometric deep learning is also increasingly being found useful in neighboring fields, such as the training of machine-learning potentials.[10,216,217] *Ab initio* calculations supported by more flexible and accurate potentials could prove immensely useful in SBDD scenarios where few or no data are available, such as in free-energy calculations on lead optimization stages. Diffusion models[218] – a family of generative models inspired by non-equilibrium thermodynamics – are also gaining increasing popularity in deep learning thanks to their generative capabilities, and have found pioneering applications in the molecular sciences.[101,127,212,219] These approaches have reached state-of-the-art in several deep learning applications and are expected to propel SBDD in the future.

A major bottleneck of AI-driven SBDD is the available training data. While protein sequence information and experimental assay data are largely available, high-quality 3D data of co-crystallized proteins and ligands with accompanying properties are largely missing. Furthermore, non- and poorly-binding molecules are often strongly under-reported in medicinal chemistry datasets. As a result, available three-dimensional datasets are often highly biased in their content,[60,161–163,220] which has historically led to poor generalizability.[61,221–223] These biases in the data are often taken advantage of by the models instead of learning true drug-target interactions.[224] Several studies have attempted to alleviate this issue, by data curation,[222,225] bias-controlled training,[121] and debiasing.[164,226] Bridging the gap between the different types of available information will be an active task in the upcoming years, with some recent work pointing in this direction already.[83]

Finally, the application of deep learning in well-established fields like chemical biology and medicinal chemistry might at times be met with skepticism by experimentalists. These well-grounded concerns commonly originate from the black-box nature of deep learning models. Additionally, robust performance benchmarks and evaluation datasets are currently missing, especially for *de novo* design studies. Despite the success of deep learning for ligand-based molecule design[17,22] and *de novo* protein design,[227] to the best of our knowledge, no deep-learning approaches for structure-based molecule discovery have been validated in the wet-lab yet. To foster broader acceptance of structure-based deep learning, we need to "open the box" and validate methods experimentally. We envision that more sophisticated applications of explainable AI[228–231] will aid in identifying underlying structure-activity relationships and binding modes and ultimately bridge the gap between theory and real-world applications.

6. Conclusions

In recent years, deep learning has taken drug discovery by storm, offering new opportunities for more efficient exploration of chemical space. Ligand binding site detection, drug-target interaction prediction, and structure-based *de novo* design can be valuable tools in early drug discovery, especially for unexplored macromolecular targets. As a whole, these approaches have great promise to extend the successes of ligand-based methods, although their full applicability in prospective scenarios has not yet been explored. Overcoming such barriers will mostly depend on additional efforts in data collection and

curation, as well as on methodologies that efficiently exploit relationships between assay and structural data.

Author Contributions

Conceptualization: all authors; Investigation: all authors; Visualization: F.G., D.v.T., R.O.; Writing - original draft: R.O., D.v.T., F.G.; Writing - review & editing: all authors.

Acknowledgements

F.G. acknowledges the support from the Centre for Living Technologies (Alliance TU/e, WUR, UU, UMC Utrecht).

Conflict of Interests

The authors declare no conflict of interests.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Keywords: artificial intelligence · *de novo* design · machine learning · medicinal chemistry · structural biology

[1] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[2] A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, M. Lackenby, G. Williamson, D. Hassabis, P. Kohli, *Nature* **2021**, *600*, 70.
[3] A. Fawzi, M. Balog, A. Huang, B. Hubert, Thomas Romera-Paredes, M. Barekatain, A. Novikov, F. J. Ruiz, J. Schrittwieser, G. Swirszcz, D. Silver, D. Hassabis, P. Kohli, *Nature* **2022**, *610*, 47.
[4] M. Ho, M. Ntampaka, M. M. Rau, M. Chen, A. Lansberry, F. Ruehle, H. Trac, *Nature Astronomy* **2022**, *6*, 936.
[5] K. Gregor, I. Danihelka, A. Graves, D. Rezende, D. Wierstra, *International Conference on Machine Learning*, PMLR **2015**, 1462–1471.
[6] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, N. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583.
[7] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. Degiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. Christopher Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, *Science* **2021**, *373*, 871.
[8] M. H. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604.
[9] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434.
[10] I. Batatia, D. P. Kovács, G. N. Simm, C. Ortner, G. Csányi, *arXiv* preprint, **2022**, arXiv:2206.07697.
[11] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, B. Kozinsky, *Nat. Commun.* **2022**, *13*, 1.
[12] J. Jiménez-Luna, F. Grisoni, N. Weskamp, G. Schneider, *Expert Opin. Drug Discovery* **2021**, *16*, 949.
[13] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, *Drug Discovery Today* **2018**, *23*, 1241.
[14] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay, J. J. Collins, *Cell* **2020**, *180*, 688.
[15] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* **2018**, *4*, 120.
[16] W. Yuan, D. Jiang, D. K. Nambiar, L. P. Liew, M. P. Hay, J. Bloomstein, P. Lu, B. Turner, Q.-T. Le, R. Tibshirani, P. Khatri, M. G. Moloney, A. C. Koong, *J. Chem. Inf. Model.* **2017**, *57*, 875.
[17] D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700153.
[18] C. M. Dobson, *Nature* **2004**, *432*, 824.
[19] G. Schneider, *Nat. Rev. Drug Discovery* **2018**, *17*, 97.
[20] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.* **2018**, *9*, 513.
[21] N. Brown, M. Fiscato, M. H. Segler, A. C. Vaucher, *J. Chem. Inf. Model.* **2019**, *59*, 1096.
[22] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R. R. Shayakhmetov, A. Zhebrak, L. I. Minaeva, B. A. Zagribelnyy, L. H. Lee, R. Soll, D. Madge, L. Xing, T. Guo, A. Aspuru-Guzik, *Nat. Biotechnol.* **2019**, *37*, 1038.
[23] A. Wlodawer, J. Vondrasek, *Ann. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 249.
[24] A. C. Anderson, *Chem. Boil.* **2003**, *10*, 787.
[25] E. E. Rutenber, R. M. Stroud, *Structure* **1996**, *4*, 1317.
[26] A. S. Reddy, S. Zhang, *Exp. Rev. Clin. Pharmacol.* **2013**, *6*, 41.
[27] Y. Kawasaki, E. Freire, *Drug Discovery Today* **2011**, *16*, 985.
[28] D. van Tilborg, A. Alenicheva, F. Grisoni, *J. Chem. Inf. Model.* **2022**, *62*, 5938.
[29] O. Civelli, R. K. Reinscheid, Y. Zhang, Z. Wang, R. Fredriksson, H. B. Schiöth, *Ann. review Pharmacol. Toxicol.* **2013**, *53*, 127.
[30] L. Kurgan, F. Miri Disfani, *Curr. Protein Pept. Sci.* **2011**, *12*, 470.
[31] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH **2009**.
[32] D. Foster, *Generative deep learning: teaching machines to paint, write, compose, and play*, O'Reilly Media **2019**.
[33] M. M. Bronstein, J. Bruna, T. Cohen, P. Velickovic, *arXiv* preprint, **2021**, arXiv:2104.13478.
[34] G. M. Morris, M. Lim-Wilby in *Molecular Modeling of Proteins*, Springer, **2008**, pp. 365–382.
[35] C. I. Branden, J. Tooze, *Introduction to Protein Structure*, Garland Science, **2012**.
[36] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, *Proc. Nat. Acad. Sci.* **2021**, *118*, e2016239118.
[37] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, *Nat. Methods* **2019**, *16*, 1315.
[38] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, *Bioinformatics* **2022**, *38*, 2102.
[39] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, B. Rost, *BMC Bioinf.* **2019**, *20*, 1.
[40] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, *arXiv* preprint, **2020**, arXiv:2007.06225.
[41] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. Vishwanathan, A. J. Smola, H.-P. Kriegel, *Bioinformatics* **2005**, *21*, i47.
[42] J. Ingraham, V. Garg, R. Barzilay, T. Jaakkola, *Adv. neural information processing systems* **2019**, *32*.
[43] S. J. Teague, *Nat. Rev. Drug Discovery* **2003**, *2*, 527.
[44] C. A. Orengo, F. M. G. Pearl, J. E. Bray, A. E. Todd, A. C. Martin, L. Lo Conte, J. M. Thornton, *Nucleic Acids Res.* **1999**, *27*, 275.
[45] L. Orellana, *Front. Mol. Biosci.* **2019**, *6*, 117.
[46] E. Di Cera, *Biophys. Rev. Lett.* **2020**, *1*, 011303.
[47] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, D. R. Koes, *J. Chem. Inf. Model.* **2017**, *57*, 942.
[48] J. Jiménez, S. Doerr, G. Martínez-Rosell, A. S. Rose, G. De Fabritiis, *Bioinformatics* **2017**, *33*, 3036.
[49] E. N. Feinberg, D. Sur, Z. Wu, B. E. Husic, H. Mai, Y. Li, S. Sun, J. Yang, B. Ramsundar, V. S. Pande, *ACS Cent. Sci.* **2018**, *4*, 1520.
[50] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl in *International Conference on Machine Learning*, PMLR, **2017**, pp. 1263–1272.

[51] X. Kong, W. Huang, Y. Liu, *arXiv* preprint, **2022**, arXiv:2208.06073.
[52] B. Lee, F. M. Richards, *J. Mol. Biol.* **1971**, *55*, 379.
[53] M. L. Connolly, *J. Appl. Crystallogr.* **1983**, *16*, 548.
[54] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
[55] K. Atz, F. Grisoni, G. Schneider, *Nature Machine Intelligence* **2021**, *3*, 1023.
[56] R. Wang, X. Fang, Y. Lu, S. Wang, *J. Med. Chem.* **2004**, *47*, 2977.
[57] K. Wüthrich, *J. Biol. Chem.* **1990**, *265*, 22059.
[58] J.-P. Renaud, A. Chari, C. Ciferri, W.-t. Liu, H.-W. Rémigy, H. Stark, C. Wiesmann, *Nat. Rev. Drug Discovery* **2018**, *17*, 471.
[59] A. M. Davis, S. J. Teague, G. J. Kleywegt, *Angew. Chem. Int. Ed.* **2003**, *42*, 2718.
[60] L. Chen, A. Cruz, S. Ramsey, C. J. Dickson, J. S. Duca, V. Hornak, D. R. Koes, T. Kurtzman, *PLoS One* **2019**, *14*, e0220113.
[61] J. Sieg, F. Flachsenberg, M. Rarey, *J. Chem. Inf. Model.* **2019**, *59*, 947.
[62] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, *bioRxiv* **2022**, 2022–07.
[63] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, *Nucleic Acids Res.* **2022**, *50*, D439.
[64] M. L. Hekkelman, I. de Vries, R. P. Joosten, A. Perrakis, *bioRxiv* **2021**.
[65] T. J. Lane, *Nat. Methods* **2023**, 1–4.
[66] S. J. Chen, M. Hassan, R. L. Jernigan, K. Jia, D. Kihara, A. Kloczkowski, S. Kotelnikov, D. Kozakov, J. Liang, A. Liwo, S. Matysiak, J. Meller, C. Micheletti, J. C. Mitchell, S. Mondal, R. Nussinov, K. Okazaki, D. Padhorny, J. Skolnick, T. R. Sosnick, G. Stan, I. Vakser, X. Zou, G. D. Rose, *Proc. Nat. Acad. Sci.* **2023**, *120*, e2214423119.
[67] A. David, S. Islam, E. Tankhilevich, M. J. Sternberg, *J. Mol. Biol.* **2022**, *434*, 167336.
[68] J. M. Thornton, R. A. Laskowski, N. Borkakoti, *Nat. Medicine* **2021**, *27*, 1666.
[69] H. Frauenfelder, S. G. Sligar, P. G. Wolynes, *Science* **1991**, *254*, 1598.
[70] R. A. Copeland, *Exp. Opin. Drug Discovery* **2010**, *5*, 305.
[71] R. Nussinov, P. G. Wolynes, *Phys. Chem. Chem. Phys.* **2014**, *16*, 6321.
[72] R. Nussinov, M. Zhang, Y. Liu, H. Jang, *J. Phys. Chem. B* **2022**, *126*, 6372.
[73] L. Klarner, M. Reutlinger, T. Schindler, C. Deane, G. Morris in *ICML 2022 2nd AI for Science Workshop*.
[74] A. Bender, I. Cortes-Ciriano, *Drug Discovery Today* **2021**, *26*, 1040.
[75] I. Cortes-Ciriano, A. Bender, T. E. Malliavin, *J. Chem. Inf. Model.* **2015**, *55*, 1413.
[76] H. Berman, K. Henrick, H. Nakamura, *Nat. Struct. Mol. Biol.* **2003**, *10*, 980.
[77] J. Desaphy, G. Bret, D. Rognan, E. Kellenberger, *Nucleic Acids Res.* **2014**, *43*, D399.
[78] J. Yang, A. Roy, Y. Zhang, *Nucleic Acids Res.* **2012**, *41*, D1096.
[79] M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin, H. A. Carlson, *Nucleic Acids Res.* **2007**, *36*, D674.
[80] R. D. Smith, J. J. Clark, A. Ahmed, Z. J. Orban, J. B. Dunbar Jr, H. A. Carlson, *J. Mol. Biol.* **2019**, *431*, 2423.
[81] M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, *J. Med. Chem.* **2012**, *55*, 6582.
[82] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, M. K. Gilson, *Nucleic Acids Res.* **2007**, *35*, D198.
[83] M. Brocidiacono, P. Francoeur, R. Aggarwal, K. Popov, D. Koes, A. Tropsha, *ChemRxiv*. **2022**.
[84] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100.
[85] I. Kufareva, A. V. Ilatovskiy, R. Abagyan, *Nucleic Acids Res.* **2012**, *40*, D535.
[86] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, T. Aittokallio, *J. Chem. Inf. Model.* **2014**, *54*, 735.
[87] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, P. P. Zarrinkar, *Nat. Biotechnol.* **2011**, *29*, 1046.
[88] J. Schmidhuber, S. Hochreiter, *Neural. Comput.* **1997**, *9*, 1735.
[89] H. Öztürk, A. Özgür, E. Ozkirimli, *Bioinformatics* **2018**, *34*, i821.
[90] M. Karimi, D. Wu, Z. Wang, Y. Shen, *Bioinformatics* **2019**, *35*, 3329.
[91] B. Shin, S. Park, K. Kang, J. C. Ho in *Machine Learning for Healthcare Conference*, PMLR **2019**, pp. 230–248.
[92] H. A. Gaspar, M. Ahmed, T. Edlich, B. Fabian, Z. Varszegi, M. Segler, J. Meyers, M. Fiscato, *ChemRxiv preprint* **2021**, 14604720v1.
[93] M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, H. Lu, *J. Proteome Res.* **2017**, *16*, 1401.
[94] Z. Cheng, Q. Zhao, Y. Li, J. Wang, *Bioinformatics* **2022**.
[95] I. Wallach, M. Dzamba, A. Heifets, *arXiv preprint* **2015**, arXiv:1510.02855.
[96] L. Zheng, J. Fan, Y. Mu, *ACS Omega* **2019**, *4*, 15956.
[97] J. Lim, S. Ryu, K. Park, Y. J. Choe, J. Ham, W. Y. Kim, *J. Chem. Inf. Model.* **2019**, *59*, 3981.
[98] H. Cho, E. K. Lee, I. S. Choi, *Sci. Rep.* **2020**, *10*, 1.
[99] H. Stärk, O. Ganea, L. Pattanaik, R. Barzilay, T. Jaakkola in *International Conference on Machine Learning*, PMLR, **2022**, pp. 20503–20521.
[100] Y. Zhang, H. Cai, C. Shi, B. Zhong, J. Tang, *arXiv preprint* **2022**, arXiv:2210.06069.
[101] G. Corso, H. Stärk, B. Jing, R. Barzilay, T. Jaakkola, *arXiv preprint* **2022**, arXiv:2210.01776.
[102] M. Tsubaki, K. Tomii, J. Sese, *Bioinformatics* **2019**, *35*, 309.
[103] V. Gligorijevic, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, R. J. Xavier, R. Knight, K. Cho, R. Bonneau, *Nat. Commun.* **2021**, *12*, 1.
[104] I. Lee, H. Nam, *J. Cheminf.* **2022**, *14*, 1.
[105] M. Jiang, Z. Li, Y. Bian, Z. Wei, *BMC Bioinf.* **2019**, *20*, 1.
[106] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, P. Siedlecki, *Sci. Rep.* **2020**, *10*, 1.
[107] J. Kandel, H. Tayara, K. T. Chong, *J. Cheminf.* **2021**, *13*, 1.
[108] J. Tubiana, D. Schneidman-Duhovny, H. J. Wolfson, *Nat. Methods* **2022**, 1–10.
[109] J. E. McGreig, H. Uri, M. Antczak, M. J. Sternberg, M. Michaelis, M. N. Wass, *Nucleic Acids Res.* **2022**, *50*, W13.
[110] A. Meller, M. D. Ward, J. H. Borowsky, J. M. Lotthammer, M. Kshirsagar, F. Oviedo, J. L. Ferres, G. Bowman, *Biophys. J.* **2023**, *122*, 445a.
[111] P. Gainza, F. Sverrisson, F. Monti, E. Rodola, D. Boscaini, M. Bronstein, B. Correia, *Nat. Methods* **2020**, *17*, 184.
[112] F. Sverrisson, J. Feydy, B. E. Correia, M. M. Bronstein in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2021**, pp. 15272–15281.
[113] S. K. Mylonas, A. Axenopoulos, P. Daras, *Bioinformatics* **2021**, *37*, 1681.
[114] D. Grechishnikova, *Sci. Rep.* **2021**, *11*, 1.
[115] G. Uludogan, E. Ozkirimli, K. O. Ulgen, N. Karalı, A. Özgür, *Bioinformatics* **2022**, *38*, ii155.
[116] Y. Zhang, S. Li, M. Xing, Q. Yuan, H. He, S. Sun, *ACS Omega* **2023**.
[117] M. Skalic, A. Varela-Rial, J. Jiménez, G. Martínez-Rosell, G. De Fabritiis, *Bioinformatics* **2019**, *35*, 243.
[118] M. Ragoza, T. Masuda, D. R. Koes, *Chem. Sci.* **2022**, *13*, 2701.
[119] Y. Li, J. Pei, L. Lai, *Chem. Sci.* **2021**, *12*, 13664.
[120] A. D. McNaughton, M. S. Bontha, C. R. Knutson, J. A. Pope, N. Kumar, *arXiv preprint* **2022**, arXiv:2205.10473.
[121] L. Chan, R. Kumar, M. Verdonk, C. Poelking, *arXiv preprint* **2022**, arXiv:2204.10663.
[122] M. Skalic, D. Sabbadin, B. Sattarov, S. Sciabola, G. De Fabritiis, *Mol. Pharm.* **2019**, *16*, 4282.
[123] H. Zhu, R. Zhou, J. Tang, M. Li, *arXiv preprint* **2022**, arXiv:2207.00821.
[124] S. R. Krishnan, N. Bung, S. R. Vangala, R. Srinivasan, G. Bulusu, A. Roy, *J. Chem. Inf. Model.* **2021**, *62*, 5100.
[125] J. Zhang, H. Chen, *J. Chem. Inf. Model.* **2022**.
[126] A. Schneuing, Y. Du, C. Harris, A. Jamasb, I. Igashov, W. Du, T. Blundell, P. Lió, C. Gomes, M. Welling, M. Bronstein, B. Correia, *arXiv preprint* **2022**, arXiv:2210.13695.
[127] I. Igashov, H. Stärk, C. Vignac, V. G. Satorras, P. Frossard, M. Welling, M. Bronstein, B. Correia, *arXiv preprint* **2022**, arXiv:2210.05274.
[128] T. Fu, W. Gao, C. W. Coley, J. Sun, *arXiv preprint* **2022**, arXiv:2211.16508.
[129] A. Ünlü, E. Çevrim, A. Sarıgün, H. Çelikbilek, H. A. Güvenilir, A. Koyaş, D. C. Kahraman, A. Rifaioglu, A. Olgaç, *arXiv preprint* **2023**, arXiv:2302.07868.
[130] S. R. Krishnan, N. Bung, S. Padhi, G. Bulusu, P. Misra, M. Pal, S. Oruganti, R. Srinivasan, A. Roy, *J. Molecular Graphics and Modelling* **2023**, *118*, 108361.
[131] M. Wang, C.-Y. Hsieh, J. Wang, D. Wang, G. Weng, C. Shen, X. Yao, Z. Bing, H. Li, D. Cao, T. Hou, *J. Med. Chem.* **2022**, *65*, 9478.
[132] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, K. Najarian, *Briefings Bioinf.* **2020**, *22*, 247.
[133] K. Abbasi, P. Razzaghi, A. Poso, M. Amanlou, J. B. Ghasemi, A. Masoudi-Nejad, *Bioinformatics* **2020**, *36*, 4633.
[134] Q. Zhao, G. Duan, M. Yang, Z. Cheng, Y. Li, J. Wang, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2022**.

[135] N. R. Monteiro, J. L. Oliveira, J. P. Arrais, *Comput. Biol. Med.* **2022**, *147*, 105772.
[136] H. Öztürk, E. Ozkirimli, A. Özgür, *arXiv preprint* **2019**, arXiv:1902.04166.
[137] R. Özçelik, H. Öztürk, A. Özgür, E. Ozkirimli, *Mol. Inf.* **2021**, *40*, 2000212.
[138] Q. Feng, E. Dueva, A. Cherkasov, M. Ester, *arXiv preprint* **2018**, arXiv:1807.09741.
[139] N. O'Boyle, A. Dalke, *ChemRxiv preprint* **2018**, 7097960v1.
[140] I. Lee, J. Keum, H. Nam, *PLoS Comput. Biol.* **2019**, *15*, e1007129.
[141] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, S. Venkatesh, *Bioinformatics* **2021**, *37*, 1140.
[142] J. Wang, N. V. Dokholyan, *J. Chem. Inf. Model.* **2022**, *62*, 463.
[143] J. Jiménez, M. Skalic, G. Martinez-Rosell, G. De Fabritiis, *J. Chem. Inf. Model.* **2018**, *58*, 287.
[144] H. Hassan-Harrirou, C. Zhang, T. Lemmin, *J. Chem. Inf. Model.* **2020**, *60*, 2791.
[145] J. Jiménez-Luna, L. Pérez-Benito, G. Martinez-Rosell, S. Sciabola, R. Torella, G. Tresadern, G. De Fabritiis, *Chem. Sci.* **2019**, *10*, 10911.
[146] A. T. McNutt, D. R. Koes, *J. Chem. Inf. Model.* **2022**, *62*, 1819.
[147] W. Torng, R. B. Altman, *J. Chem. Inf. Model.* **2019**, *59*, 4131.
[148] Q. Liu, P.-S. Wang, C. Zhu, B. B. Gaines, T. Zhu, J. Bi, M. Song, *J. Mol. Graphics Modell.* **2021**, *105*, 107865.
[149] V. R. Somnath, C. Bunne, A. Krause, *Advances in Neural Information Processing Systems* **2021**, *34*, 25244.
[150] Z. Qiao, W. Nie, A. Vahdat, T. F. Miller III, A. Anandkumar, *arXiv preprint* **2022**, arXiv:2209.15171.
[151] J. C. Pereira, E. R. Caffarena, C. N. Dos Santos, *J. Chem. Inf. Model.* **2016**, *56*, 2495.
[152] F. Gentile, V. Agrawal, M. Hsing, A.-T. Ton, F. Ban, U. Norinder, M. E. Gleave, A. Cherkasov, *ACS Cent. Sci.* **2020**, *6*, 939.
[153] X. Wang, G. Terashi, C. W. Christoffer, M. Zhu, D. Kihara, *Bioinformatics* **2020**, *36*, 2113.
[154] A. T. McNutt, P. Francoeur, R. Aggarwal, T. Masuda, R. Meli, M. Ragoza, J. Sunseri, D. R. Koes, *J. Cheminf.* **2021**, *13*, 1.
[155] H. Zhang, L. Liao, K. M. Saravanan, P. Yin, Y. Wei, *PeerJ* **2019**, *7*, e7362.
[156] J. A. Morrone, J. K. Weber, T. Huynh, H. Luo, W. D. Cornell, *J. Chem. Inf. Model.* **2020**, *60*, 4170.
[157] O. Méndez-Lucio, M. Ahmad, E. A. del Rio-Chanona, J. K. Wegner, *Nature Machine Intelligence* **2021**, *3*, 1033.
[158] K. A. Stafford, B. M. Anderson, J. Sorenson, H. van den Bedem, *J. Chem. Inf. Model.* **2022**, *62*, 1178.
[159] M. Masters, A. H. Mahmoud, Y. Wei, M. A. Lill in *ICLR2022 Machine Learning for Drug Discovery*, **2022**.
[160] M. Volkov, J.-A. Turk, N. Drizard, N. Martin, B. Hoffmann, Y. Gaston-Mathé, D. Rognan, *J. Med. Chem.* **2022**, 7946–7958 .
[161] K. Peng, Z. Obradovic, S. Vucetic in *Biocomputing 2004*, World Scientific, **2003**, pp. 435–446.
[162] L. Chaput, J. Martinez-Sanz, N. Saettel, L. Mouawad, *J. Cheminf.* **2016**, *8*, 1.
[163] V.-K. Tran-Nguyen, C. Jacquemard, D. Rognan, *J. Chem. Inf. Model.* **2020**, *60*, 4263.
[164] R. Özçelik, A. Bag, B. Atıl, M. Barsbey, A. Özgür, E. Özkırımlı, *arXiv preprint* **2022** arXiv:2107.05556.
[165] H. Zhu, J. Yang, N. Huang, *J. Chem. Inf. Model.* **2022**, pMID: 36268980.
[166] J. Zaucha, C. A. Softley, M. Sattler, D. Frishman, G. M. Popowicz, *Chem. Commun.* **2020**, *56*, 15454.
[167] S. Park, C. Seok, *J. Chem. Inf. Model.* **2022**, *62*, 3157, pMID:35749367.
[168] S. Pérot, O. Sperandio, M. A. Miteva, A.-C. Camproux, B. O. Villoutreix, *Drug Discovery Today* **2010**, *15*, 656.
[169] J. Zhao, Y. Cao, L. Zhang, *Comput. Struct. Biotechnol. J.* **2020**, *18*, 417.
[170] R. A. Laskowski, *J. Mol. Graphics* **1995**, *13*, 323.
[171] G. P. Brady, P. F. Stouten, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 383.
[172] J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, T. A. Funkhouser, *PLoS Comput. Biol.* **2009**, *5*, e1000585.
[173] M. Weisel, E. Proschak, G. Schneider, *Chemistry Central Journal* **2007**, *1*, 1.
[174] Y. Cui, Q. Dong, D. Hong, X. Wang, *BMC Bioinf.* **2019**, *20*, 1.
[175] S. H. Khan, H. Tayara, K. T. Chong, *Cells* **2022**, *11*, 2117.
[176] I. Kozlovskii, P. Popov, *Communications biology* **2020**, *3*, 1.
[177] I. Kozlovskii, P. Popov, *J. Chem. Inf. Model.* **2021**, *61*, 3814.
[178] R. Krivák, D. Hoksza, *J. Cheminf.* **2018**, *10*, 1.
[179] D. G. Isom, C. A. Castañeda, B. R. Cannon, B. García-Moreno, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 5260.
[180] S. Chakravarty, R. Varadarajan, *Structure* **1999**, *7*, 723.
[181] L. Gagliardi, A. Raffo, U. Fugacci, S. Biasotti, W. Rocchia, H. Huang, B. B. Amor, Y. Fang, Y. Zhang, X. Wang, C. Christoffer, D. Kihara, A. Axenopoulos, S. Mylonas, P. Daras, *Computers & Graphics* **2022**, *107*, 20.
[182] D. Ni, Z. Chai, Y. Wang, M. Li, Z. Yu, Y. Liu, S. Lu, J. Zhang, *WIREs Comput. Mol. Sci.* **2022**, *12*, e1585.
[183] P. Schneider, G. Schneider, *J. Med. Chem.* **2016**, *59*, 4077.
[184] P. Ertl, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374.
[185] C. Lipinski, A. Hopkins, *Nature* **2004**, *432*, 855.
[186] M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, *J. Cheminf.* **2017**, *9*, 1.
[187] R. V. Devi, S. S. Sathya, M. S. Coumar, *Applied Soft Computing* **2015**, *27*, 543.
[188] D. Douguet, E. Thoreau, G. Grassy, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 449.
[189] L. G. Ferreira, R. N. Dos Santos, G. Oliva, A. D. Andricopulo, *Molecules* **2015**, *20*, 13384.
[190] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, P. Battaglia, *arXiv preprint* **2018** arXiv:1803.03324.
[191] B. Samanta, A. De, G. Jana, V. Gómez, P. K. Chattaraj, N. Ganguly, M. Gomez-Rodriguez, *Journal of machine learning research* **2020** *Apr. 21 (114)*, 1–33.
[192] N. De Cao, T. Kipf, *arXiv preprint* **2018**, arXiv:1805.11973.
[193] M. Popova, O. Isayev, A. Tropsha, *Sci. Adv.* **2018**, *4*, eaap7885.
[194] N. Ståhl, G. Falkman, A. Karlsson, G. Mathiason, J. Bostrom, *J. Chem. Inf. Model.* **2019**, *59*, 3166.
[195] A. Gupta, A. T. Müller, B. J. Huisman, J. A. Fuchs, P. Schneider, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700111.
[196] M. Moret, L. Friedrich, F. Grisoni, D. Merk, G. Schneider, *Nature Machine Intelligence* **2020**, *2*, 171.
[197] M. Thomas, A. Bender, C. de Graaf, *Curr. Opin. Struct. Biol.* **2023**, *79*, 102559.
[198] M. Filipavicius, M. Manica, J. Cadow, M. R. Martinez, *arXiv preprint* **2020**, arXiv:2012.03084.
[199] S. Chithrananda, G. Grand, B. Ramsundar, *arXiv preprint* **2020**, arXiv:2010.09885.
[200] A. Powers, H. Yu, P. Suriana, R. Dror, *bioRxiv* **2022**.
[201] D. Flam-Shepherd, K. Zhu, A. Aspuru-Guzik, *Nat. Commun.* **2022**, *13*, 1.
[202] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Advances in neural information processing systems* **2014**, *27*.
[203] F. Grisoni, B. J. H. Huisman, A. L. Button, M. Moret, K. Atz, D. Merk, G. Schneider, *Sci. Adv.* **2021**, *7*, eabg3338.
[204] M. Moret, M. Helmstädter, F. Grisoni, G. Schneider, D. Merk, *Angew. Chem. Int. Ed.* **2021**, *60*, 19477.
[205] N. S. Hatzakis, *Biophys. Chem.* **2014**, *186*, 46.
[206] H.-B. Guo, A. Perminov, S. Bekele, G. Kedziora, S. Farajollahi, V. Varaljay, K. Hinkle, V. Molinero, K. Meister, C. Hung, P. Dennis, N. Kelley-Loughnane, R. Berry, *Sci. Rep.* **2022**, *12*, 10696.
[207] Y. Min, Y. Wei, P. Wang, N. Wu, S. Bauer, S. Zheng, Y. Shi, Y. Wang, X. Wang, D. Zhao, J. Wu, J. Zeng, *arXiv preprint* **2022**, arXiv:2208.10230.
[208] G. M. Lee, C. S. Craik, *Science* **2009**, *324*, 213.
[209] A. Valente, C. Miyamoto, F. L Almeida, *Curr. Med. Chem.* **2006**, *13*, 3697.
[210] K. Gunasekaran, B. Ma, R. Nussinov, *Proteins Struct. Funct. Bioinf.* **2004**, *57*, 433.
[211] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, *IEEE Signal Processing Magazine* **2017**, *34*, 18.
[212] E. Hoogeboom, V. G. Satorras, C. Vignac, M. Welling in *International Conference on Machine Learning*, PMLR, **2022**, 8867–8887.
[213] S. Zhang, Y. Liu, L. Xie, *arXiv preprint* **2022**, arXiv:2206.02789.
[214] V. R. Somnath, M. Pariset, Y.-P. Hsieh, M. R. Martinez, A. Krause, C. Bunne, *arXiv preprint* **2023**, arXiv:2302.11419.
[215] P. Gainza, S. Wehrle, A. Van Hall-Beauvais, A. Marchand, A. Scheck, Z. Harteveld, D. Ni, S. Tan, F. Sverrisson, C. Goverde, P. Turelli, C. Raclot, A. Teslenko, M. Pacesa, S. Rosset, S. Georgeon, J. Marsden, A. Petruzzella, K. Liu, Z. Xu, Y. Chai, P. Han, G. F. Gao, E. Oricchio, B. Fierz, D. Trono, H. Stahlberg, M. Bronstein, B. E. Correia, *bioRxiv* **2022**.
[216] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, B. Kozinsky, *Nat. Commun.* **2022**, *13*, 2453.
[217] R. Galvelis, A. Varela-Rial, S. Doerr, R. Fino, P. Eastman, T. E. Markland, J. D. Chodera, G. De Fabritiis, *arXiv preprint* **2022**, arXiv:2201.08110.
[218] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli in *International Conference on Machine Learning*, PMLR, **2015**, pp. 2256–2265.
[219] J. Yim, B. L. Trippe, V. De Bortoli, E. Mathieu, A. Doucet, R. Barzilay, T. Jaakkola, *arXiv preprint* **2023**, arXiv:2302.02277.
[220] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, P. Netrapalli, *Advances in Neural Information Processing Systems* **2020**, *33*, 9573.
[221] F. Boyles, C. M. Deane, G. M. Morris, *Bioinformatics* **2020**, *36*, 758.

[222] I. Wallach, A. Heifets, *J. Chem. Inf. Model.* **2018**, *58*, 916.
[223] A. Gonczarek, J. M. Tomczak, S. Zareba, J. Kaczmar, P. Dabrowski, M. J. Walczak, *Comput. Biol. Med.* **2018**, *100*, 253.
[224] J. Yang, C. Shen, N. Huang, *Front. Pharmacol.* **2020**, *11*, 69.
[225] J. Scantlebury, N. Brown, F. Von Delft, C. M. Deane, *J. Chem. Inf. Model.* **2020**, *60*, 3722.
[226] V. Sundar, L. Colwell, *J. Chem. Inf. Model.* **2019**, *60*, 56.
[227] A. H.-W. Yeh, C. Norn, Y. Kipnis, D. Tischer, S. J. Pellock, D. Evans, P. Ma, G. R. Lee, J. Z. Zhang, I. Anishchenko, B. Coventry, L. Cao, J. Dauparas, S. Halabiya, M. DeWitt, L. Carter, K. N. Houk, D. Baker, *Nature* **2023**, *614*, 774.
[228] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, in *Machine Learning and Knowledge Extraction* (Eds.: A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl), Springer International, Cham **2018**, pp. 295–303.
[229] J. Jiménez-Luna, F. Grisoni, G. Schneider, *Nature Machine Intelligence* **2020**, *2*, 573.
[230] J. Tan, Y. Zhang, *arXiv preprint* **2023**, DOI: arXiv:2301.11765.
[231] T. H. Vo, N. T. K. Nguyen, Q. H. Kha, N. Q. K. Le, *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2112–2123.

Manuscript received: December 24, 2022
Revised manuscript received: March 30, 2023
Accepted manuscript online: April 4, 2023
Version of record online: June 13, 2023