Binding Affinity Prediction with 3D Machine Learning: Training Data and Challenging External Testing

Jose Carlos Gomez Tamayo [a], *Lili Cao* [a], *Mazen Ahmad* [a], *Gary Tresadern* [a, *]

[a] In Silico Discovery, Janssen Research & Development, Janssen Pharmaceutica N. V.,
Turnhoutseweg 30, B-2340 Beerse, Belgium.

* Corresponding author: E-mail: gtresade@its.jnj.com. Tel: +32 1464 1569

Keywords: 3D ML, absolute affinity prediction, CNN

ABSTRACT

Protein-ligand binding affinity prediction is one of the major challenges in computational assisted drug discovery. An active area of research uses machine learning (ML) models trained on 3D structures of protein ligand complexes to predict binding modes, discriminate active and inactives, or predict affinity. Methodological advances in deep learning, and artificial intelligence along with increased experimental data (3D structures and bioactivities) has led to many studies using different architectures, representation, and features. Unfortunately, many models do not learn details of interactions or the underlying physics that drive protein-ligand affinity, but instead just memorize patterns in the available training data with poor generalizability and future use. In this work we incorporate "dense", feature rich datasets that contain up to several thousand analogue molecules per drug discovery target. For the training set, PDBbind dataset is used with enrichment from 8 internal lead optimization (LO) datasets and inactive and decoy poses in a variety of combinations. A variety of different model architectures was used and the model performance was validated using the binding affinity for 12 internal LO and 6 ChEMBL external test sets. Results show a significant improvement in the performance and generalization power, especially for virtual screening and suggest promise for the future of ML protein-ligand affinity prediction with a greater emphasis on training using datasets that capture the rich details of the affinity landscape.

INTRODUCTION

Structure-based virtual screening (SBVS) is an important tool for early drug discovery. Classically, SBVS is performed using physics-based methods such as docking in which the 3D coordinates of a hypothetical ligand and the target are first identified and then scored according to an empirical scoring-function.[1] The ligand placement is a strength of docking methods, and in many cases, they can retrieve molecules in correct binding poses. The more problematic aspect is the ranking or scoring of the ligands and poses based on predicted binding affinity. A recent data driven approach based on diffusion generative modeling outperformed docking at identifying bound poses in a blind docking task.[2] As well as ligand positioning and scoring, the study also required binding site identification, a problem for which docking is not intended and for which other methods are better suited[3,4]. This may partly account for the worse performance of docking in that study, but the 'all-in-one' diffusion model was an impressive solution compared to protocols of multistep and multimethod physics-based approaches. Thus, the key ligand scoring step in docking is known to underperform compared to placement and suffers from a lack of sampling and the underlying oversimplicity of docking scoring-functions. More powerful and expensive physics-based ensemble approaches such as free energy perturbation can be used for this task.[5–7] However, these methods have a high computational cost, technical difficulty, and sensitivity to input setup, that hampers large scale virtual screening and is the subject of active research.[8–11]

The breakthroughs in machine learning (ML) and deep learning (DL) approaches coupled with the increasing experimental structures and bioactivity data have led to the emergence of 3D ML and DL models as an alternative to docking scoring functions.[12,13] The 3D coordinates of protein-ligand complexes are used as input but the aims for specific models can be varied: predict binding poses, separate binders from non-binders, rank based on activity or predict activity. The goals are attractive not only because of the value for drug discovery, but also the need for improvement versus docking, particularly in high throughput applications. There are various caveats for most reported ML/DL affinity prediction models that are usually true for docking. They arise from the simplification of ligand partitioning from solvent to protein-bound phases and include the lack of an ensemble approach, the absence of (de)solvation terms, the simplified or absence of entropic considerations, etc. Learning affinity from single experimental 3D protein-ligand complexes emphasizes the interaction potential between protein and ligand and only indirectly captures entropy or (de)solvation factors.

DL models for binding affinity prediction were reported using different architectures and representations.[14–23] In brief, several key approaches to featurization have been pursued. One approach uses protein-ligand interaction fingerprints that are defined using the atomic interactions or contacts between protein and ligand that fall within a specific radius of the ligand[24]. An alternative method uses voxels that encode atomic properties at 3D grid-points in the protein-ligand binding site[14,15,25] and combine with a 3D convolutional neural network (3DCNN). Finally, graph descriptions of the protein-ligand complex combined with graph convolutional neural networks (3DGNN),[21,23] have proved efficient in comparison to 3D convolutions due to their 3D invariant encoding. Nevertheless, the increasing complexity of description and model architectures has not led to a significant improvement in performance. The latest described 3DGNNs do not significantly outperform the first generation of CNNs and more importantly none of the models reported show general predictivity power for new targets.[25–32]

The PDBbind dataset[33] is widely used for training affinity prediction models. It (v.2020) contains approximately 20000 3D protein–ligand X-ray structures of known binding affinity measured as the inhibition constant (K_i), dissociation constant (K_d), or half-maximum inhibition concentrations (IC_{50}). The set is further split into a refined set of around 5000 high-quality X-ray structures and the most reliable affinity data (K_i and K_d only) and a core set of 290 complexes from 58 different proteins with five different ligands of various affinities. Either with or without data augmentation (inactive compounds, docking pose augmentation) it represents the best public data source to train 3D models. However, various concerns regarding ML/DL affinity prediction have arisen and the composition and completeness of the training dataset is at the forefront.[34–36] The major problem relates to the sparsity of the protein-ligand training matrix.[35]

The sparsity of the training dataset is the fraction of the ligand-protein matrix having a missing experimental binding affinity. A dataset with a large representation of targets can cover more of the available protein space but the sparsity of the data is high, as usually few of the ligands are tested against the biological space and the dataset becomes heavily biased. Simple experiments have shown that 3D ML/DL affinity models based only on protein description or only on ligand description (omitting interactions) can show modest to good performance when trained and tested on PDBbind data.[35–38] In other words, there are biases in the data that are relatively easy to uncover, the bioactivity is likely clustered for each protein and chemical series due to the lack of sufficient protein-ligand complexes per target.[35] In contrast, drug discovery involves exploration of structure activity relationships (SAR) of hundreds or thousands of similar molecules in the same protein binding site. There is a richness and subtlety to a dense bioactivity landscape, it contains implicit redundancy over a large descriptor space and number of dimensions. For example, activity cliffs are small changes in ligand structure leading to unexpected outcomes in bioactivity and are relevant in drug discovery and can be traced to the 3D description [39,40]. We hypothesize that it is necessary to learn from highly dense and redundant protein-ligand datasets to understand the rugged 3D landscape of bioactivity and identify the meaningful origins of activity versus memorizable patterns in the data. Accordingly, ML/DL affinity prediction models trained on more dense training data with more generalizable aims will need alternative and improved external testing, closer to real world drug discovery scenarios.

Without an organized effort, the data sparsity problem is unlikely to be resolved by the evolution of experimental crystallography which is costly and usually avoids redundancy, focusing on novel protein and ligands, further increasing the sparsity problem. Although there has been no robust quantification, we speculate that the current number of 15-20K experimental structures needs to increase by orders of magnitude to rival the size of datasets used in other common ML supported drug-design tasks. It seems computational modeling will be required to reach this level of data and one possible solution is to use molecular docking of known actives, but we must avoid adding noise from incorrect docked binding poses. For some active molecules the uncertainty can be reduced if it has exceptionally high chemical similarity to the ligand in an experimental protein ligand structure. Fortunately, the lead optimization (LO) stage in drug discovery generates such datasets because it involves the synthesis of many analogue compounds with occasional examples being co-crystalized with the protein target. Therefore, we here generate high confidence ligand-protein complexes by using maximum common substructure docking (MCS) for series of compounds with high similarity to experimentally crystalized ligands. We demonstrate that the decrease of data sparsity increases model performance and generalization power. Also, we discuss the use of pose decoys to force the model to learn from the interactions thus avoiding ligand overfitting and, training one part of the network on ligand efficiency to force the model to learn from the interaction type and not by the number of interactions which in turn is correlated to molecular weight, a common bias leading to overfitting on the compound size.

METHODS

Datasets. For initial training of our models and comparison of performance with previous work we begin by using the widely employed PDBbind dataset. Then, we increase the training data by including drug discovery LO datasets derived from different protein targets. We also use similar datasets built from data retrieved from ChEMBL[41] The details are shown below, but these datasets were chosen to be denser and have a larger chemical coverage and typical affinity range. Ligands were selected to be similar to those in the resolved experimental 3D protein ligand structures. Dataset details:

- PDBbind v.2019: Comprises 17679 protein-ligand complexes, each accompanied by experimentally determined inhibition constants, including K_i, dissociation constants (K_d), and IC_{50}.
- PDBbind v.2019 Refined: A subset of the PDBbind v.2019 dataset, consisting of 4852 complexes that have been subjected to stricter quality criteria.[42] The final number of structures after curation is 4543 excluding the core set 2016, typically used for testing. The overall range of affinity is from 2.0 to 11.9 log units. The mean range in affinity for a specific protein target is 2.2 log units when considering proteins co-crystalized with at least two ligands.

- PDBbind v.2019 Non-refined: A distinct subset that incorporates an additional 4719 complexes (4629 after data curation) sourced from the general set, but which are not present in the refined subset [43]. The overall range of affinity is from 2.0 to 12 log units for all data, the mean range in affinity for a specific protein target is 1.8 log units.
- PDBbind 2016 core set: A representative subset of PDBbind refined dataset widely used for benchmarking purposes. From the initial 285 complexes compiling this dataset, 260 survived data curation. Affinities in this dataset fall between 2.1 and 11.8, while the mean range in affinity for a specific protein target is 4.6 log units.
- In house drug discovery LO datasets: Twenty protein targets and associated ligands corresponding to previous structure-enabled drug discovery projects at Janssen were chosen. Eight target datasets were used for training and the remaining were used as new, challenging, external test sets, Table 1 (training) and Table 2 (external test sets). An unusually larger proportion of datasets were used for external testing in this study as we are particularly interested to inspect the model performance in real-world scenarios, and the training data is still substantially augmented versus PDBbind alone. The number of experimental structures used for each of the protein targets ranged from 5 to 443, with a total of 811 and 1170 for the training and test sets. The total number of complexes in the training and testing datasets was 16318 and 27381 respectively. Thus, on average the inclusion of the docked close analogues led to a factor of 20-23 increase versus using only experimentally solved structures. The minimum and maximum number of protein-ligand complexes per target across the datasets was 384 and 3995. The overall range of affinity for all the training set data was from 3.1 to 11.8 and the mean range of affinity per target was 5.4 log units. We note how the total set of LO drug discovery data had a narrower overall range compared to PDBbind (8.7 vs 10 log

units in total), but each given target had a wider range of 5.4 log units compared to ~2 log units in PDBbind.

- ChEMBL LO datasets: We examined the human protein targets in ChEMBL with the most bioactivity data and chose known drug discovery proteins (avoiding ADMET related proteins for instance) for which experimental 3D protein structures were available. Crystal structures were then retrieved for these proteins from the protein databank[44] filtering to retain only X-ray structures with 2.5 Å or better resolution. Compounds with experimental data were extracted from the ChEMBL database[41] for the specified proteins. The pChEMBL value was used as the activity measure. Overall, the 6 datasets were prepared using docking to augment the size of the dataset as described later. Unlike the prior LO datasets, the increased chemical diversity in the ChEMBL data did not always allow for reliable MCS docking. Instead, we performed constrained docking by specifying simple and well-known key interaction(s) that need to be satisfied, such as the hinge binding motif for the kinase targets, or the interaction with the catalytic aspartates in beta secretase. The details for the datasets are summarized in Table 3. The overall range of affinity is from 2 to 11.2 log units for all data, the mean range in affinity for a specific protein target is 7.0 log units. The mean number of experimental structures per target was 32, and the mean number of final complexes per target was 2218. Thus, the density of the data was increased compared to the experimental structures alone, and denser than the PDBbind data but not as dense as internal drug discovery LO datasets.

Table 1. Details of 3D drug discovery LO datasets used for model training. Number of active compounds successfully docked with the MCS protocol across 8 proprietary datasets used for model training.

Dataset	Total 3D P-L complexes	Crystal structures	Min affinity	Max affinity	Activity range
ROS1	384	6	5.0	8.3	3.3
hOGA	1499	9	3.1	9.5	6.4
FGFR1	2533	443	5.0	11.8	6.8
BACE1	2410	106	5.0	9.3	4.3
PRMT5	582	43	3.9	9.9	6.0
HPK1	2417	26	5.0	11.3	6.3
MALT1	2498	48	4.5	8.2	3.7
IL17	3995	130	4.3	8.9	4.6

Table 2. Details of 3D drug discovery in house LO datasets used for external model testing. Total active compounds docked, number of active compounds successfully docked with MCS, number of inactive compounds docked, and total number of 3D protein-ligand complexes built for each target.

Dataset	nActives	nInactive s	Total	X-rays	Min affinity	Max affinity	Activity range
FGFR2	247	1713	2310	5	5.6	9.5	3.9
PIK3G	772	11979	13199	128	4.1	10.0	5.9
CGAS	807	34390	35740	179	4.3	9.1	4.7
PIK3D	624	24457	25710	81	4.5	10.2	5.7
CSNK1D	754	29646	31117	165	3.7	9.3	5.6
MENIN	1388	46172	48510	124	4.6	10.9	6.3
KDR	557	21227	22356	82	4.6	8.8	4.2
PDE2	1401	47339	49849	17	3.3	10.0	6.7
NIK	146	4539	6002	24	5.0	9.2	4.2
MGL	797	27747	29169	128	4.3	11.1	6.8
PDE10	1420	48246	50725	21	3.8	9.5	5.7
MCL1	472	20886	21650	216	2.7	9.5	6.9

Table 3. Details of the ChEMBL LO datasets used for external model testing. Total active compounds docked, number of active compounds successfully docked through MCS, number of inactive compounds docked, and total number of 3D protein-ligand complexes built for each target.

Dataset	nActives	nInactives	Total	X-rays	Min affinity	Max affinity	Activity range
EGFR	3220	27604	30824	36	3.9	11.2	7.3
ABL	1090	10338	11428	17	4.0	10.8	6.8
MAPK	1349	60094	61443	50	4.4	11.0	6.6
PLK1	428	7125	7553	6	3.2	9.1	5.9
VEGF2	2846	53654	56500	21	3.7	10.7	7.0
BACE1	4026	17100	21478	60	2.5	11.0	8.4

Data Augmentation: Actives. Recognizing data density to be a critical challenge we augmented the dataset as follows:

Firstly, internal projects were ranked by the number of available crystal structures. Success of the MCS docking heavily depends on the number of experimental structures and their scaffold diversity. Maximizing the number of available structures guarantees a higher number of matches from the compounds with available bioassay data. Secondly, for each selected target, biochemical assays containing pX_{50} (inhibition or binding) were extracted from our internal database. The criteria for selecting assays include the number of tested compounds with a non-qualified pIC_{50} value, the range of activity and the shape of the activity distribution, preferring those assays with a good representation of high affinity compounds, which are usually less common. Additionally, assays for the same target showing high correlation ($R^2 > 0.8$) and an average error below the experimental error (< ~0.5 pIC_{50} units) were combined for the sake of maximum data retrieval.

Extracted compounds for each target were thus docked against the available X-ray structures using MCS docking within the GLIDE molecular docking software.[45] The settings used included

SP (standard precision) docking, snap MCS core constraints with a tolerance of 0.3 Å and shape similarity maximization. Since the same compound is docked against different structures, we keep the resulting complex structure that contains the higher number of atoms in the MCS to the crystallized ligand. If this criterion does not retrieve a unique structure, then the docking with the better docking score is selected. Finally, the docked compounds were filtered according to the number of atoms and shape similarity. The procedure involves setting a minimum size for the scaffold used in the MCS docking and a minimum shape similarity. Setting an MCS minimum size excludes unrealistic dockings based on small fragment MCS matches such as single aromatic rings, while shape similarity reduces the possibility of docked compounds diverging from the reference. For each target, these values were heuristically selected based on iterative visual inspection of the results.

For ChEMBL datasets we used a similar approach but found it necessary to use feature constrained docking instead of MCS docking. Our aim is to increase the size of the datasets but not at any cost, instead we want to only retain high confidence docked solutions that remain close in 3D to experimental solutions. In brief, the ChEMBL data comprises examples of LO series from medicinal chemistry literature but there are fewer actives per scaffold than compared with a full internal LO dataset. Also, similar but distinct chemical scaffolds from different publications can be seen to share binding modes but not necessarily a substantial MCS. Consider for example a change in a nitrogen position in an aromatic scaffold of a kinase inhibitor, wherein similar substituents and the hinge binding motif are maintained despite no substantial MCS. Thus, upon testing we found that docking with interaction restraints was far better at returning the expected overlapping and high confidence docking results. For each target we identified interactions that are most repeated and well stablished for the given protein (i.e: kinase hinge hydrogen bond acceptor and donor/s). The sequence motifs in which the interactions are located were used to identify the residues taking part in the interactions across the different PDB structures as residue and atom numbers are usually not consistent. Constrained docking was then performed over all available structures, keeping the complex with better docking score per ligand docking solution, visual inspection was performed to remove remaining problematic docked structures. Finally, macrocycles were identified and removed from the dataset we well as compounds with poor docking scores. The docking score threshold for each dataset was selected based on the distribution of scores, thereby removing those at the tail of worse score.

Data Augmentation: Pose Decoys. A concern with 3D ML protein-ligand binding affinity prediction models is the over-reliance on ligand information[36], which can result in models that exhibit comparable predictive capabilities irrespective of the protein structure. To mitigate this form of overfitting, one potential approach is to incorporate pose decoys into the training data[46]. These decoys, which are active molecules docked outside the binding site, or active molecules docked in just incorrect poses, possess the same ligand information but dissimilar interaction information. We assign an affinity label based on the docking score, which is normalized between 0 and 3, thereby providing the label with a degree of physical significance and facilitating the regression task done in logarithmic units of potency. In this manner, the models now train on the same ligand bound to the same protein but in different complexes, with different interactions and different assigned affinities. The models are compelled to utilize the full combination of protein and ligand information for accurate predictions.

Data Augmentation: Inactive compounds. Prior research has shown that incorporating inactive compounds during the training of DL models can enhance model performance and generalizability[21,47]. In the literature, two primary strategies have been identified: the first entails cross-docking all PDBbind compounds and assigning low affinity scores to the resulting virtual complexes[17], while the second employs decoy compounds sourced from databases such as DUD-E[48] but results show this can be problematic[34] and improvements have been suggested[49]. These decoys are presumed to be inactive due to their topological dissimilarities from the reference compounds, yet they possess analogous physicochemical characteristics. For every drug discovery LO target, we have docked confirmed bioinactive compounds selected from their respective high throughput screening data. Additionally, we have reduced the inactive compound pool by filtering for drug-like properties, thus impeding the models from readily discerning differences between inactive and active compounds. Since we dock the same compound to multiple crystal structures for each target, we keep the docking solution with the best score. We noticed that our 2DCNN model did not benefit from using this data for training, keeping the generated inactive compounds for benchmarking.

Protein and Ligand Preparation. X-ray complexes were prepared using the Schrödinger Protein Preparation Wizard.[50] In cases where residues with absent side-chain atoms were detected, they were filled using Prime.[51] PROPKA[52] was used to define the protonation states at pH 7, except for BACE1, which were determined at pH 5 matching the acidic pH used in the bioassays.

Preparation of ligands was conducted through the standardization of compounds using LigPrep tool from the Schrödinger package, featuring a single conformer and single default tautomer generation, and EPIK for generating protonation states[53]

Ligand Docking. MCS docking used the GLIDE software that requires a preprepared docking grid that was calculated using default settings specifying the crystallized ligands as the grid center. The active compounds were docked with the MCS approach, incorporating standard precision (SP) docking, snap MCS core constraints with a tolerance of 0.3 Å, and shape similarity maximization. Any compounds that were unable to dock through MCS were subsequently docked without any constraints and labeled accordingly. Similarly, inactive compounds were also docked with the same grids and the GLIDE SP approach but without any constraints. As mentioned above, the single best scored pose for an inactive versus all the structures for a given target was retained. Decoy pose generation for active compounds was accomplished via SMINA[54], as GLIDE was unable to systematically provide decoy poses. Docking scores are used as a comparison metric to assess the relative performance of the affinity prediction models versus a classical SBVS approach. In those cases, only actives and inactive molecules were considered (not decoys) and the GLIDE SP scores were used for the best pose of each ligand docked versus all the structures for a given target (i.e., the score for the specific docked solution that was used as model input and as defined by the protocol above). Altogether, we have designed a strategy to confidently augment data density taking advantage of the LO project data from Janssen. Moreover, we generated pose decoys to force the model to learn from interactions and generated challenging inactive datasets for each target following the same protocol.

Molecular Descriptors. Our 2DCNN model incorporates protein interaction fingerprints (PLEC) from the open drug discovery toolkit package (ODDT) and RDKit molecular properties[55] as inputs.

PLEC is equivalent to the extended connectivity fingerprint (ECFP)[56] which is typically used for small molecules but extended to include nearby atoms from the protein structure. The atoms included in PLEC are defined by a distance threshold, and both the ligand and receptor radii are modifiable parameters. In this work, we have used different configurations of PLEC to improve the generalizability of the resulting model as described in the model details section.

Performance Metrics.

Pearson correlation coefficient: describes the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges from -1 to +1. A value of +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and a value of 0 indicates no correlation.

Spearman ranking coefficient (ρ): describes the monotonic relationship between two continuous or ordinal variables. It is denoted by the symbol "ρ" (rho) and ranges from -1 to +1. The Spearman correlation coefficient does not assume a linear relationship between the variables. Instead, it assesses the degree to which the variables tend to increase or decrease together, regardless of whether this relationship is linear or not. A value of +1 indicates a perfect positive monotonic correlation, -1 indicates a perfect negative monotonic correlation, and 0 indicates no monotonic correlation.

Mean absolute error (MAE): is a metric used to measure the average magnitude of errors between predicted and real values.

Enrichment factor (EF): metric utilized in virtual screening studies to assess the effectiveness of a scoring function in discriminating active from inactive compounds. It is defined as the ratio of the percentage of active compounds detected in a certain fraction of the total database to the percentage of that fraction of the database. This metric gives a qualitative indication of how enriched a subset of compounds is with active molecules compared to random selection.

Fingerprint 2D similarity calculations were used to compare the diversity of molecules retrieved from different models. In this case RDKit fingerprints were used with the Tanimoto similarity metric to calculate the average similarity within sets of molecules.

2D convolutional neural network (2DCNN) Model. The 2DCNN model employed a dual architecture, consisting of two separate 2D convolutional blocks that were fed with different protein-ligand environment configurations (PLECs). Each block was paired with two multi-perceptron layers (MPL) that were used to analyze the properties of the ligand. The outputs of the two 2D blocks were concatenated with their corresponding MPL outputs and then passed through a final MPL to produce the final output. To enhance the generalization of the model, one of the blocks used a simple PLEC configuration with a 3.5 Å distance threshold and radii of 1 for both the protein and ligand. Meanwhile, the other block uses a standard configuration with a 4.5 Å distance threshold and radii of 2 and 4 for the ligand and receptor respectively. This approach allows the first block to focus on close, simple pair-to-pair interactions captured with short fingerprints. This encourages the model to describe ligand and receptor interactions by the collection of small features rather than pattern recognition of elaborate lengthy fingerprints that tend to be complex specific and lead to over training or less generalizability.

During training, the model objective is set as the sum of the outputs of both blocks, with the more generalizable block driving predictivity through a more heavily weighted loss. At each training iteration, both real and decoy poses are fed through the model, with an independent loss calculated for each. Additionally, the objective is set to the pseudo ligand efficiency (affinity divided by molecular weight) to ensure that the model focuses on the interactions rather than the number of interactions. During prediction, the affinity is recovered by multiplying it back by the molecular weight.

Additional architectures

We implemented the architectures published by Moon et al.[21] These architectures include a 3DCNN, GNN and the physics informed graph neural network PIGNet. We modified the code to allow additional data input (LO datasets). We also modify the training process by splitting the loss in two terms: public and LO losses. The losses were scaled to account for the difference in the dynamic range found in LO datasets in comparison to PDBbind. By normalizing the loss, we ensure that the different sources of data are equally weighted during the backpropagation process independently of their affinity range. These models were trained with the PDBbind dataset as well as augmented data, including docked decoys (292518), inactive complexes generated using IBS molecules (831885) and cross-docked inactive compounds generated from PDBbind (527682). We tested the published weights trained with this data plus the models trained including also internal data. The models are described below.

3D convolutional neural network (3DCNN) Model. 3DCNNs are a type of DL architecture designed to extract and learn complex features from volumetric data. Unlike 2DCNNs, which are designed to work with 2D images, 3DCNNs are specifically engineered to work with 3D data, such as CT scans, MRI images, and 3D microscopy data.[57] 3DCNNs convolve a set of filters or kernels over the entire input volume, allowing them to extract and learn features that are sensitive to the spatial and temporal relationships between the input voxels. These learned features are then used to classify, segment, or perform other tasks on the input data. This architecture has been utilized for binding affinity prediction, with kDEEP being a prominent example.[14] However, it is computationally intensive and non-equivariant, requiring grid rotations to achieve invariant predictions for different grid orientations. We implemented the version published by Moon et al.[21] This architecture is identical to kDEEP in all aspects except for the atom features, which include atom type, period, group, degree of atom, hybridization, formal charge, and aromaticity.

Graph convolutional neural networks (3DGNN). 3DGNNs are a type of neural network that can operate on non-Euclidean data structures, such as graphs. 3DGNNs have been demonstrated to outperform other architectures in tasks such as node classification, link prediction, and graph classification in both accuracy and speed.[58] In the realm of biomolecules, 3DGNNs have demonstrated superior performance in predicting molecular properties compared to other architectures.[59] Graph-based models are particularly effective due to their invariance or equivariance to complex orientation when properly defined, making them more efficient than other models such as 3DCNNs. Additionally, 3DGNNs require less computational power than 3DCNNs. Typically, the input to a 3DGNN is a graph representation of the molecule, where atoms are represented as nodes and bonds as edges. This graph-based representation enables end-to-end training, where feature extraction and prediction steps are learned concurrently from the data. The implementation uses the same atom features used in the 3DCNN to encode voxel information. The model involves several gated graph attention modules and interaction networks, with two adjacency matrices accounting for covalent bonds and intermolecular interactions.

PIGNet. The previously published PIGNet incorporates physics-based knowledge to improve generalization in binding affinity prediction.[21] The model parameterizes equations describing intermolecular interactions, such as van der Waals, hydrogen bond, metal-ligand, and hydrophobic interactions. The predicted affinity is the sum of the atom-atom pairwise interactions, divided by a penalty rotor term accounting for the loss of entropy experienced by the ligand upon complexation. By integrating physics-based knowledge with DL, PIGNet aims to improve the generalizability, accuracy, and interpretability of binding affinity predictions. The architecture is identical to the GNN described above except for the output layers. While the GNN provides the final output through fully connected layers, PIGNet feeds the resulting node features into physics-informed parameterized equations.

RESULTS

Training Test Set Performance.

Models were subject to the typical training routine in which the performance was evaluated with a test set removed from the training set for each target. Details of the model training data composition are available at Supplementary table 1. The averaged correlation metrics showed modest correlation (Table 4). However, some caution is needed because the averaged values were adversely affected by one or two very poor correlations. Supplementary table 2 shows the correlation metrics for the disaggregated training test sets. Due to the strategies employed to avoid overfitting, such as the simplified descriptors and using decoys etc. (see Methods), modest performance in these test sets was expected. Interestingly, the 2DCNN showed a noticeable improvement in correlation and errors when using internal data and a superior performance versus the rest of the models. The 3DCNN and PIGNet showed no major differences when trained with or without the additional internal data, whereas the absolute error in prediction for the 3DGNN improved substantially when trained with additional internal data. Also noteworthy, the GLIDE score correlation is similar to the majority of the DL models. Overall, the initial results from the training set performance did not show substantial improvements versus docking, except for the 2DCNN trained using additional internal data. A key interest here was to investigate the performance on unseen external test sets and therefore the generalizability of the models, and whether they learn interactions and features that describe 3D protein ligand bioactivity.

Table 4. Performance metrics for the prediction of activity for the internal validation sets for the different models trained with and without drug discovery LO data.

Architecture	r	Spearman	MAE	LO training data
2DCNN	0.19	0.15	1.05	no
2DCNN	0.56	0.54	0.84	yes
3DCNN	0.33	0.29	2.03	no
3DCNN	0.34	0.32	1.99	yes
3DGNN	0.23	0.20	4.56	no
3DGNN	0.35	0.32	1.26	yes
PIGNet	0.34	0.29	2.54	no
PIGNet	0.33	0.29	3.55	yes
GLIDE	0.33	0.29	-	-

r, Spearman and MAE are averages over the same metrics computed for the eight test sets extracted from the training data.

PDBbind Core Set Testing. Here we used the PDBbind core set as an external test of the performance of the different models. The PDBbind dataset is widely used as a standard benchmark for assessing the performance of ML and DL models in predicting binding affinity and evaluating docking tools and empirical scoring functions. However, as mentioned above, previous studies have identified potential biases in this dataset that can lead to overoptimistic performance. Although most studies refer to this dataset as an external test set, it is far from a real-world test to find new actives versus a new protein target. The PDBbind coreset is stratified from the general set. Every target present in the set is also present in the training set, so there are no new target proteins. This may explain why reports using only PDBbind as training and this external test set show modest performance when trained on protein alone. Regarding the compounds, approximately 30% are duplicates with the training set but bind to different proteins in the test set. If the activity happens to be similar between training and test sets then an overtrained model can make a correct prediction based on patterns in the chemical structure alone.

We trained our models with and without incorporating the dense LO datasets as above and tested their performance on this test set using various metrics. Firstly, we were able to reproduce the published performance values for all the architectures using their published weights except for PIGNet, whose performance was worse (correlation coefficient of 0.67 vs 0.75), Table 5. We found that including the LO datasets in training improved the performance of the models across all architectures, except for the 3DCNN which exhibited a slight decrease in performance. This improvement, of approximately 0.11 units in Pearson correlation coefficient for the 3DGNN model for example, suggests a positive impact of adding the dense new datasets to the model training. The PIGNet and 3DGNN models top the performance table with almost identical performance (r 0.79). The 2DCNN model shows the biggest improvement with the dense LO data with a Pearson r increasing from 0.5 to 0.72. This dependence can be attributed to the fact that the 2DCNN model parameters and hyperparameters were optimized using LO datasets, as such the parameters may not be ideal for PDBbind core data alone. To our surprise the performance on this test set (Table 5) was better than seen in the internal validation discussed previously (Table 4).

Table 5. Performance metrics for the prediction of activity for the PDBbind core set for the different models trained with and without drug discovery LO data. Performance metrics include Pearson correlation coefficient (r). Spearman ranking coefficient, squared Pearson correlation coefficient, coefficient of determination (R^2) and mean average error (MAE).

Architecture	**r**	**Spearman**	**MAE**	**LO training data**
2DCNN	0.50	0.50	3.15	no
2DCNN	0.72	0.7	1.9	yes
3DCNN	0.69	0.68	1.76	no
3DCNN	0.67	0.66	1.8	yes
3DGNN	0.68	0.68	2.08	no
3DGNN	0.79	0.78	1.43	yes
PIGNet	0.67	0.66	2.04	no
PIGNet	0.79	0.77	1.48	yes

Drug Discovery LO External Test Sets. Up to this point we have built models using state of the art approaches and using many more protein ligand complexes than previously reported. The models displayed similar performance on the PDBbind core set suggesting they were behaving as expected, and the additional data led to further improvement. Therefore, we set a difficult but realistic challenge to understand if such models can be used for new drug discovery targets. We deliberately chose a large number of test cases to avoid bias and aid analysis. Twelve drug discovery LO datasets were not used for training and instead reserved for external testing. Each dataset corresponds to a unique protein target that overall cover several typical drug discovery relevant protein families. Models trained with and without the drug discovery LO data training sets were tested against these target sets.

Figure 1 and table 6 show the performance of the different model architectures trained with and without internal data. Overall, the performance is poor, with low correlation coefficients in the range of 0.0 to 0.29 but in most cases a small improvement in performance is seen for those models trained with internal data. This improvement is observed for the 3DGNN and 2DCNN neural network architectures. On the other hand, PIGNet and the 3DCNN models are not benefiting from using internal data, rather the opposite. Importantly, none of the models built from the different architectures outperform GLIDE molecular docking software that shows the best average correlation.

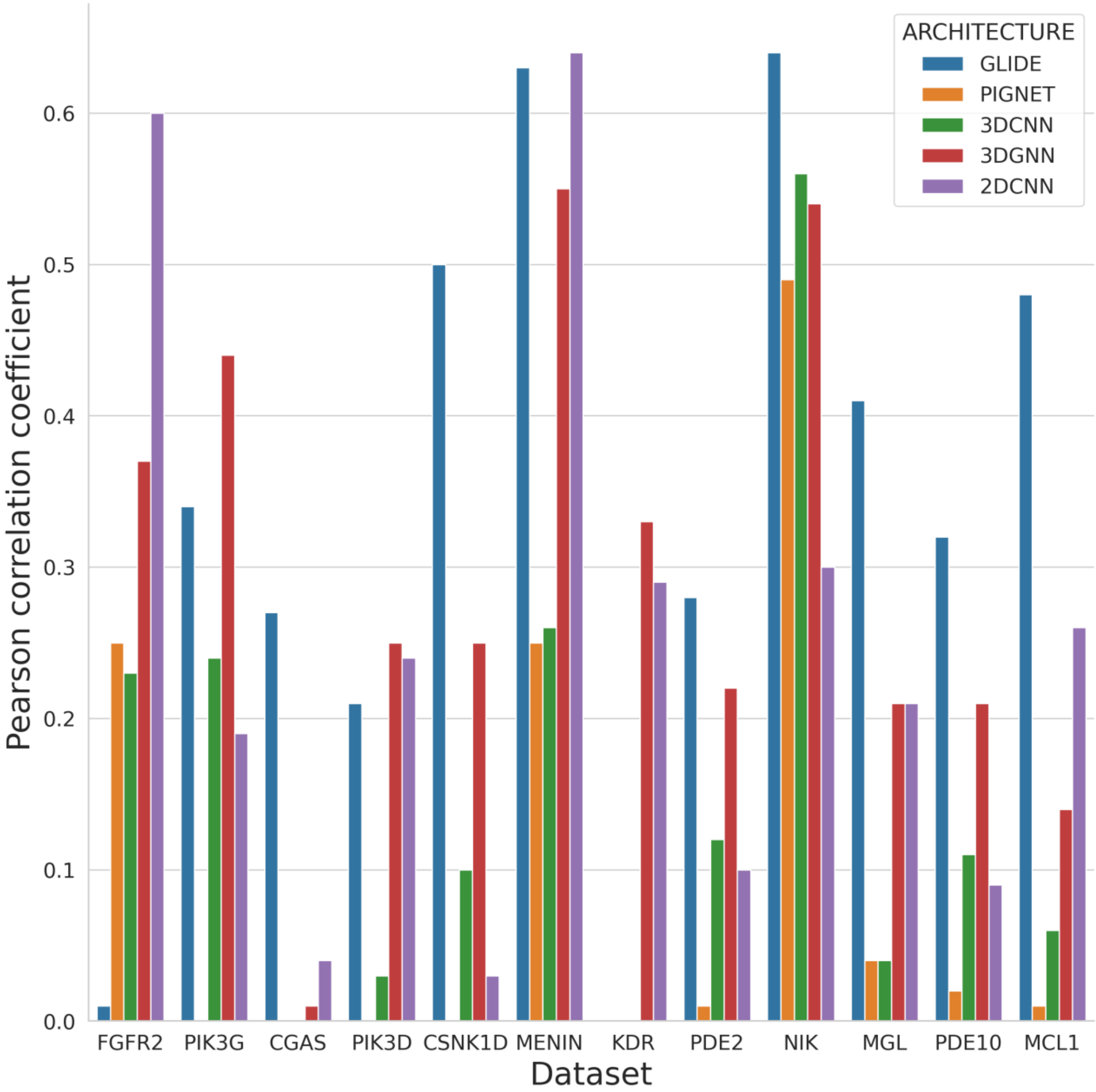

Figure 1. Pearson correlation coefficient of the different architectures per LO test set.

Table 6. Average Pearson correlation coefficient for the twelve drug discovery LO test datasets, calculated considering only active compounds.

Architecture	r	Spearman	MAE	LO training data
2DCNN	0.17	0.25	1.85	no
2DCNN	0.25	0.22	3.04	yes
3DCNN	0.19	0.15	2.32	no
3DCNN	0.12	0.13	2.58	yes
3DGNN	0.05	0.07	3.24	no
3DGNN	0.29	0.24	3.46	yes
PIGNet	0.00	-0.05	4.43	no
PIGNet	0.06	0.06	4.97	yes
GLIDE	0.34	0.22	10.58	-

r, Spearman and MAE are averages over the same metrics computed for the 12 LO test sets.

Based on our experience with models of this type we anticipated poor results to predict the absolute binding affinity for a set of entirely new protein targets. Afterall, the models treat the binding process in a highly simplified manner. However, despite the poor correlation metrics, there is an improvement in performance when using the additional LO data for training. Figure 2 shows a few examples where the 3DGNN trained with internal data showed superior performance compared to the model trained without internal LO data and GLIDE molecular docking.

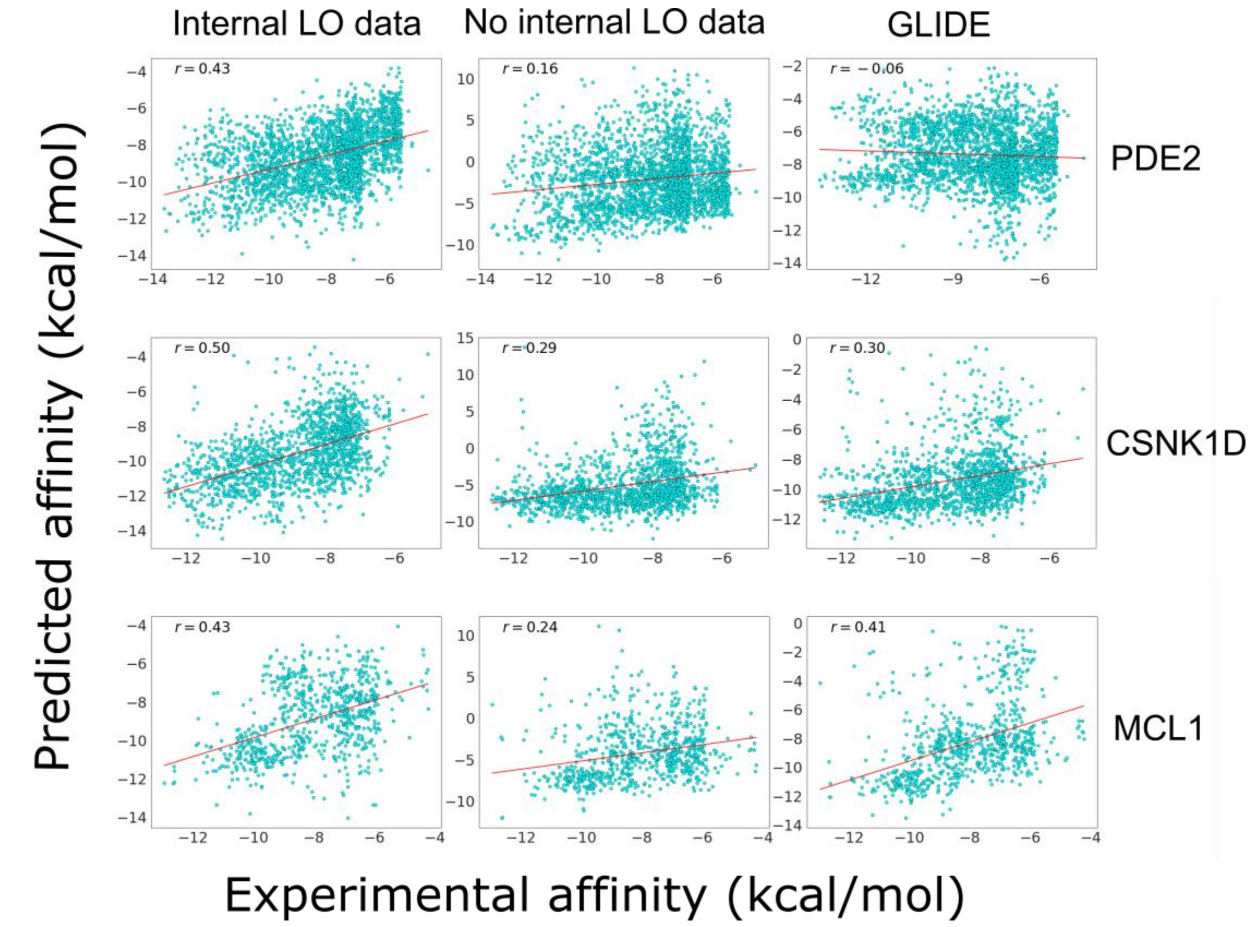

Figure 2. Experimental binding affinity versus predicted binding affinity for 3 of the LO test sets showing improved correlation when trained with internal LO data compared to the same model trained without LO data and GLIDE molecular docking software. Note the correlation in these plots is calculated using only samples with measured experimental affinity.

We therefore sought to understand if the models could still be used in a virtual screening scenario, asking the question: are they enriching the selection of more active compounds amongst the better ranked molecules? For this we calculated the enrichment factor (EF) to assess whether the models still show discriminating power against inactive compounds. Given that this is one of the main aims of this study, and we are principally interested to learn about 3D ML affinity prediction as a competitive computational tool, we again included the GLIDE docking software performance.

For the external LO datasets, we observe that the addition of drug discovery LO datasets in training improved the performance of 3DGNN and 2DCNN architectures while PIGNet and 3DCNN models showed reduced performance compared to the original models trained only with PDBbind data (Table 7). The 3DGNN model showed the best performance with an average EF at 1% of 9.0 followed by 2DCNN with 7.3 and GLIDE with 6.6. However, performance of the models is not homogenous among the different datasets as shown in Figure 3. The 3DGNN appears to be the most consistent method by ranking best in six of the twelve datasets, second in five and third in one of the datasets (Table 7). Interestingly, GLIDE performed admirably and was best for three of the LO target datasets while being last or penultimate ranked for five of the targets. Particularly, these targets represent examples of structures with challenging docking conditions. For example, the binding site of PDE2 and PDE10 contain two metallic cations and KDR shows flexibility across the structures in one of the loops forming the binding site.

Table 7. Average enrichment factor performance metric for the retrieval of actives for the twelve drug discovery LO test datasets. The enrichment factor is the number of active compounds retrieved in top 1% predictions compared to random and averaged for each model architecture over all the drug discovery target datasets. Models were trained with and without LO datasets.

Architecture	**EF_1%**	**LO training data**
2DCNN	3.1	no
2DCNN	7.3	yes
3DCNN	6.3	no
3DCNN	3.1	yes

3DGNN	6.3	no
3DGNN	9.0	yes
PIGNet	2.6	no
PIGNet	2.0	yes
GLIDE	6.6	-

EF_1% is the average enrichment factor at top 1% predictions for the 12 LO test sets.

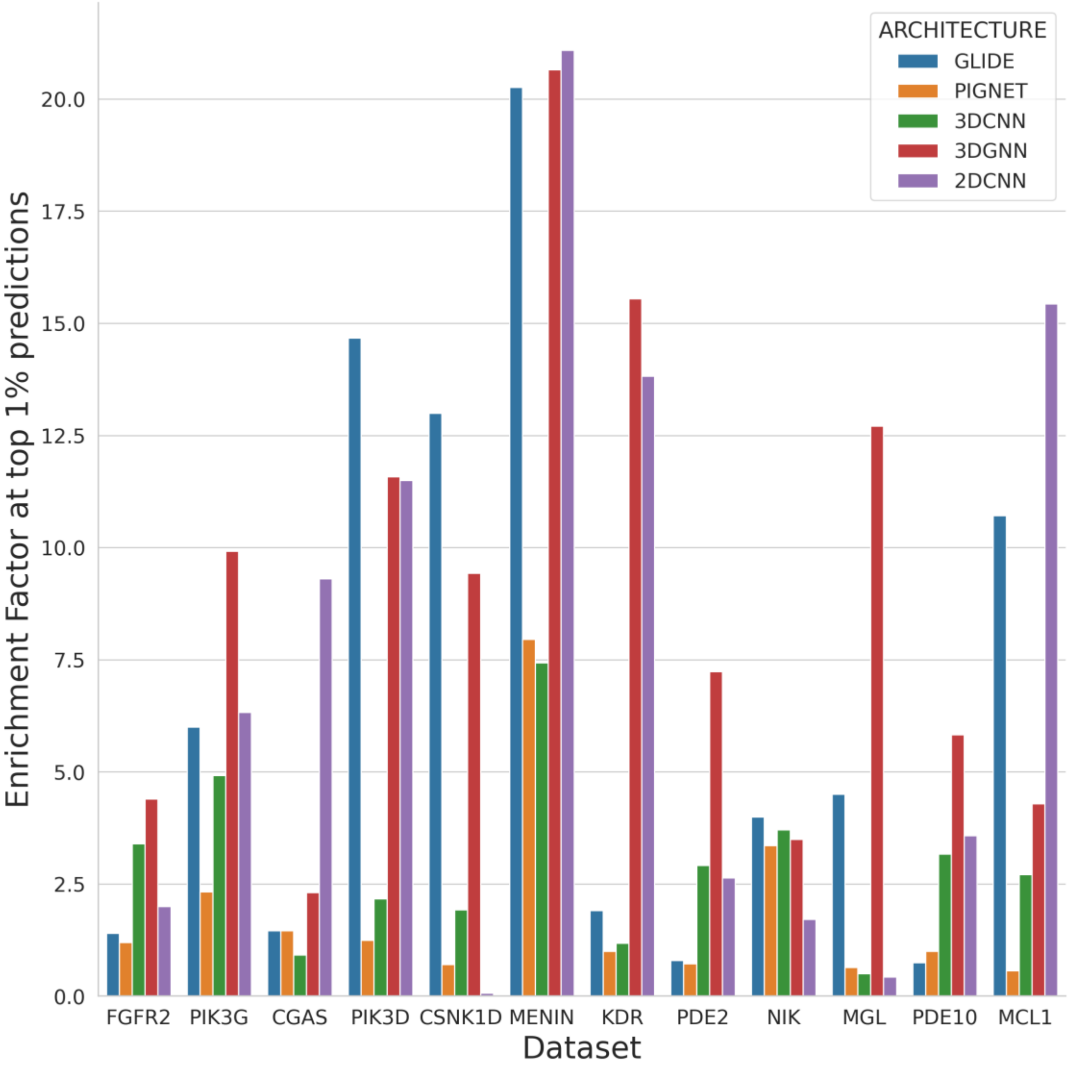

Figure 3. The enrichment factor for the retrieval of active compounds in the top 1% of the predictions of the different architectures per LO test set. Data shown for models trained with the LO datasets.

MW bias. One critical issue we considered is the molecular weight bias. MW is a simple property that can show a low positive correlation with affinity and models tend to shortcut learning by overweighting the importance of MW or a correlated feature (number of atoms, number of interactions). To know if our models trained with the drug discovery LO data are providing enrichment in different MW ranges, we stratified the data by MW and computed the enrichment factor. Surprisingly, our models did not show a bias in enrichment towards high MW ranges. Instead, the best enrichments are found within the range 350 to 500 Daltons. Figure 4 shows the average top 1% EF by method for all datasets. Noteworthy, 3DGNN is showing consistently good EFs (top 1%) in all MW ranges except for the 500-550 range where GLIDE is better.

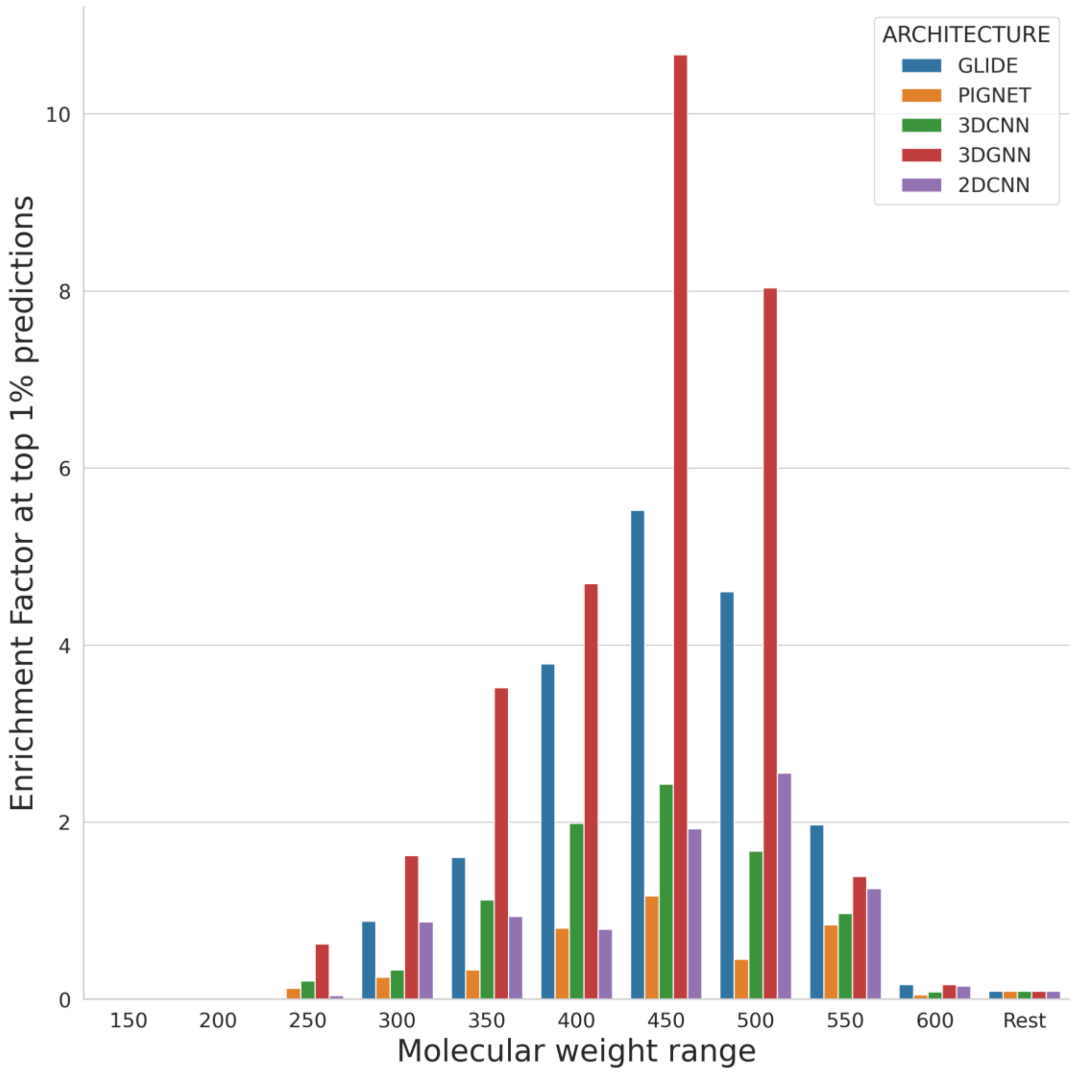

Figure 4. Active compounds enrichment factor at the top 1% predictions for various MW ranges using different architectures on the external LO test datasets. The MW ranges are divided into bins separated by 50 Daltons, and each tick on the x-axis represents the range of its value minus 50. Data shown for models trained with the LO datasets.

Hit diversity analysis. We investigated whether the models were predicting the bioactivity of chemically diverse or similar molecules in comparison to GLIDE as a representative state of the art docking software. We computed the average similarity among the top 100 active molecules (real active compounds) ranked by each method (Table 8). The results show that the 3DGNN returned the most similar actives with an average Tanimoto similarity of 0.6, while PIGNet produced the most diverse set with an average similarity of 0.49. However, these differences were dependent on the dataset, as the differences were smaller between models within most of the datasets (Figure 5).

Table 8. Comparing the average 2D fingerprint similarity of the top 100 predicted actives per model architecture.

ARCHITECTURE	Average similarity
2DCNN	0.57
3DCNN	0.51
3DGNN	0.60
PIGNet	0.49
GLIDE	0.53

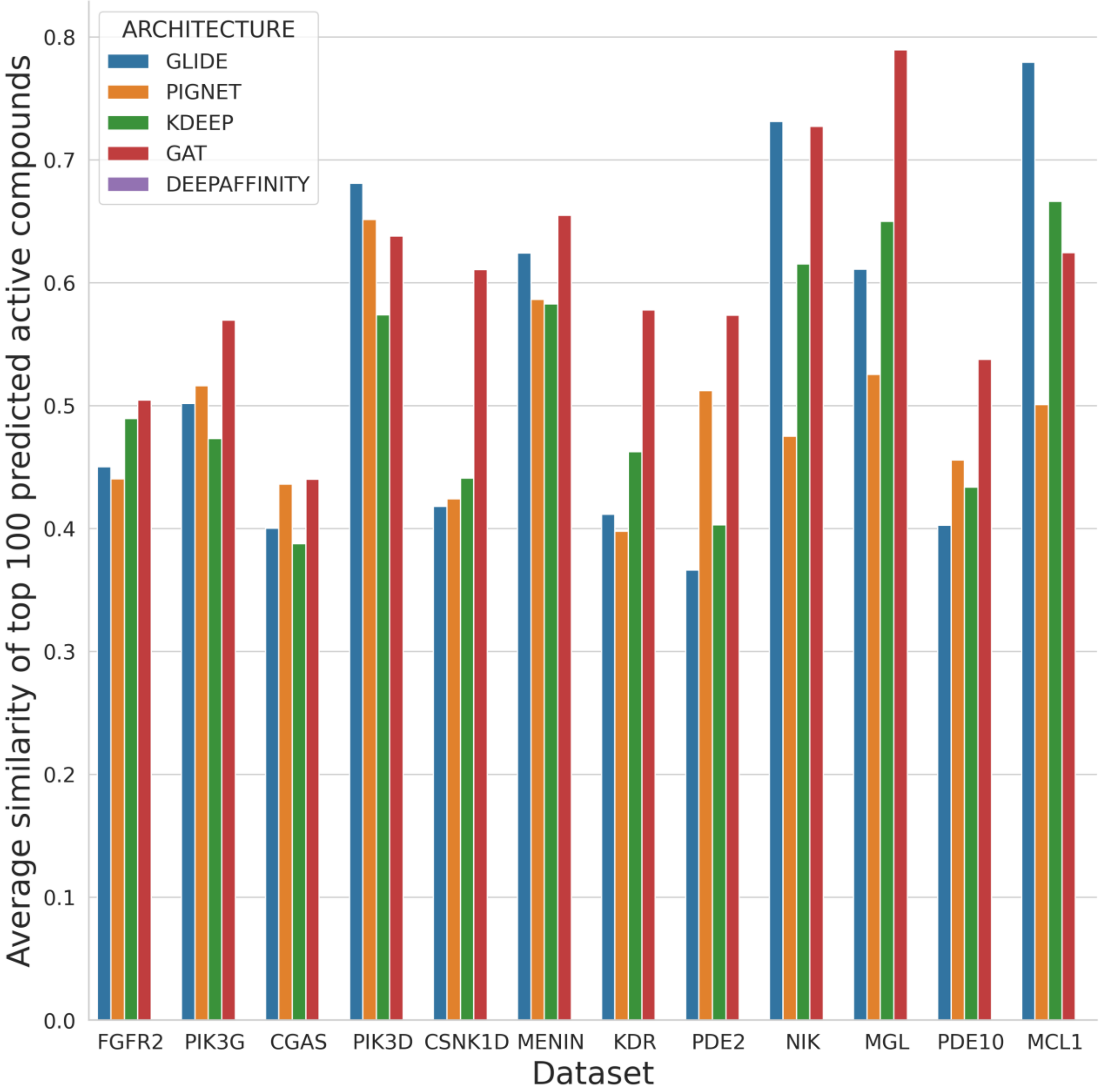

Figure 5. Average cross-similarity for the top predicted 100 active compounds per LO external test set and model architecture. Data shown for models trained with the LO datasets.

External ChEMBL LO test sets. To add further diversity to the test sets, and consider data available in the public domain, we assessed the performance of our models on prepared ChEMBL datasets, as described in the Methods, both with and without the drug discovery LO training data. Again, the inclusion of the LO data resulted in improved performance for all models except for the 3DCNN (Table 9). Notably, the 3DGNN model trained with internal LO data showed the best performance in both Pearson correlation coefficient and enrichment in the top 1% with averages of 0.5 and 10.70 respectively. GLIDE appears as the second best in terms of correlation coefficient (r 0.47), while the 3DGNN trained without the LO data ranked second in terms of enrichment factor (9.53).

Table 9. ChEMBL datasets average Pearson correlation coefficient (r), Spearman correlation coefficient, mean average error (MAE) and enrichment factor of at top 1% (EF_1%) predictions among the different neural network architectures used in this work with and without drug discovery LO data in training.

Architecture	r	Spearman	MAE	EF_1%	Internal data
2DCNN	0.21	0.28	2.02	2.20	no
2DCNN	0.17	0.05	1.29	7.19	yes
3DCNN	0.31	0.28	2.68	7.46	no
3DCNN	0.11	0.10	1.69	3.38	yes
3DGNN	0.30	0.34	2.81	9.53	no
3DGNN	0.50	0.37	1.38	10.65	yes
PIGNet	0.07	0.04	3.31	3.25	no
PIGNet	0.16	0.13	3.51	9.49	yes
GLIDE	0.47	0.38	10.60	7.99	no

r, Spearman, MAE, and EF_1% are averages over the same metrics computed for the 6 ChEMBL test sets.

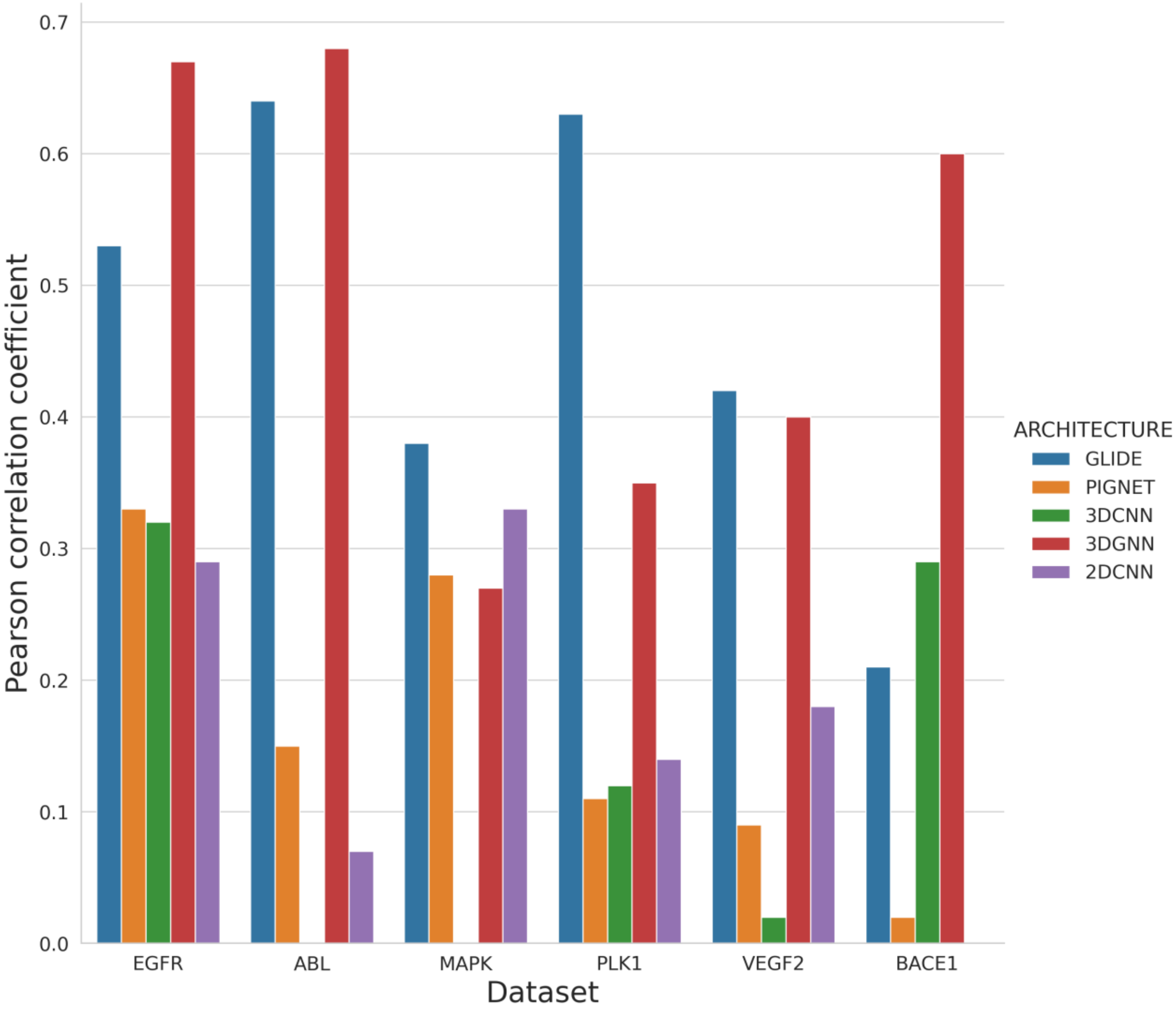

Figure 6. Pearson correlation coefficient of the different architectures per ChEMBL test set trained with LO data.

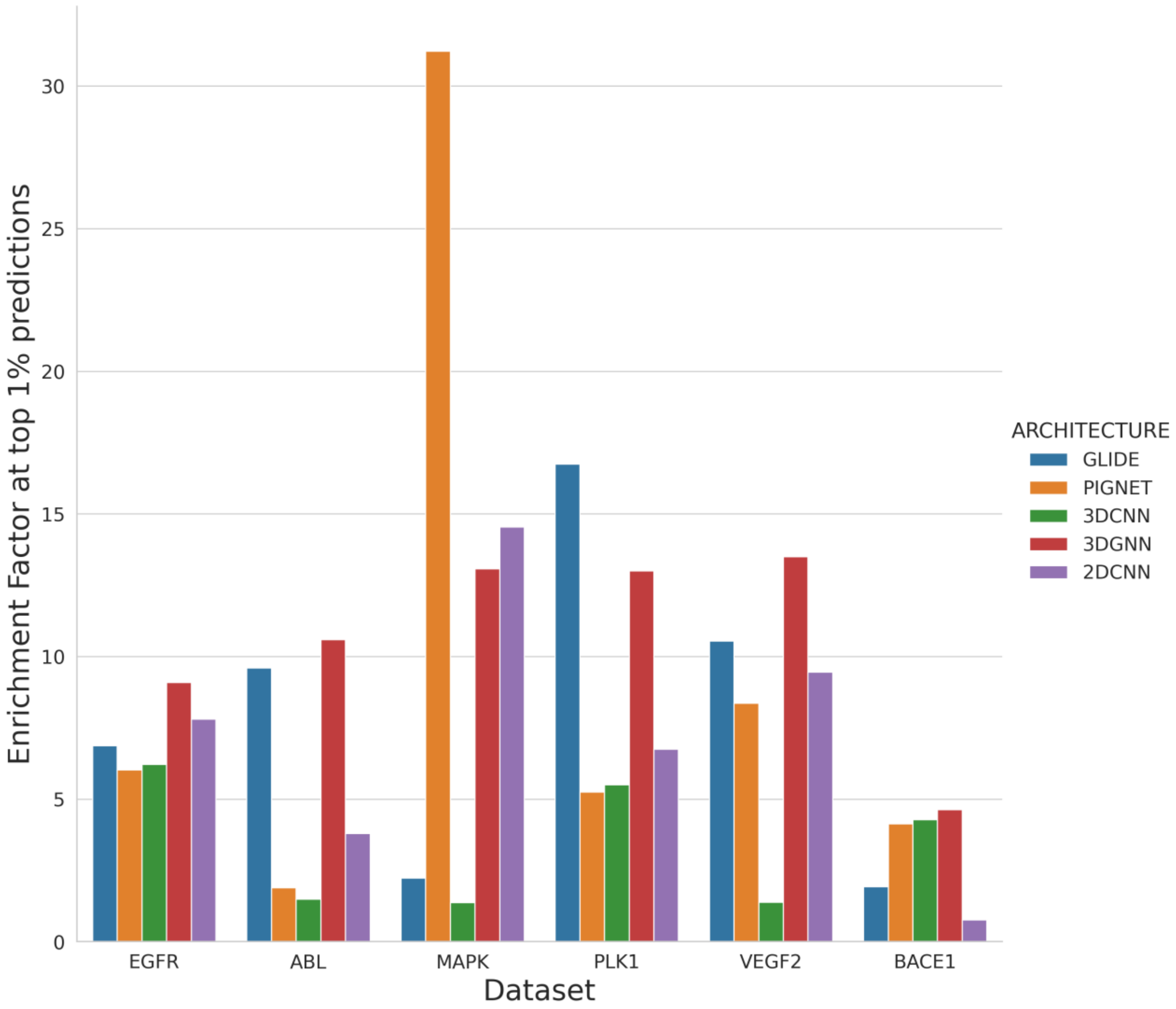

Figure 7. The enrichment factor for the retrieval of active compounds in the top 1% of the predictions of the different architectures per ChEMBL test set trained with LO data.

The performance by dataset and neural network architecture is shown in Figures 6 (Pearson correlation coefficient) and 7 (enrichment factor) for the models trained with internal data. The 3DGNN shows the most consistent performance in both correlation and enrichment factor, ranking first in 3 and 4 out of the 6 datasets respectively. GLIDE ranks first in 3 out of 6 datasets in terms of Pearson correlation coefficient. However, GLIDE falls to the 4th position in average enrichment factor (7.99), a noticeable drop in performance compared to the 3DGNN trained with internal data.

The 3DGNN model trained without internal data ranked second in average enrichment factor (9.53), but with a noticeable drop in Pearson correlation coefficient (0.30). Interestingly, the PIGNet model trained with internal data achieved 3rd position in terms of enrichment factor, with very low correlation coefficient. However, average enrichment factor for this model is biased by the disproportionate performance in a single dataset (MAPK, Figure 7). The 2DCNN architecture ranked 5th in average enrichment factor (7.19), showing on the other hand low average correlation coefficient (0.17). Lastly, the 3DCNN architecture, the only one which showed no improvement by training with internal LO data, showed an average enrichment factor of 7.49 and an average Pearson correlation coefficient of 0.31.

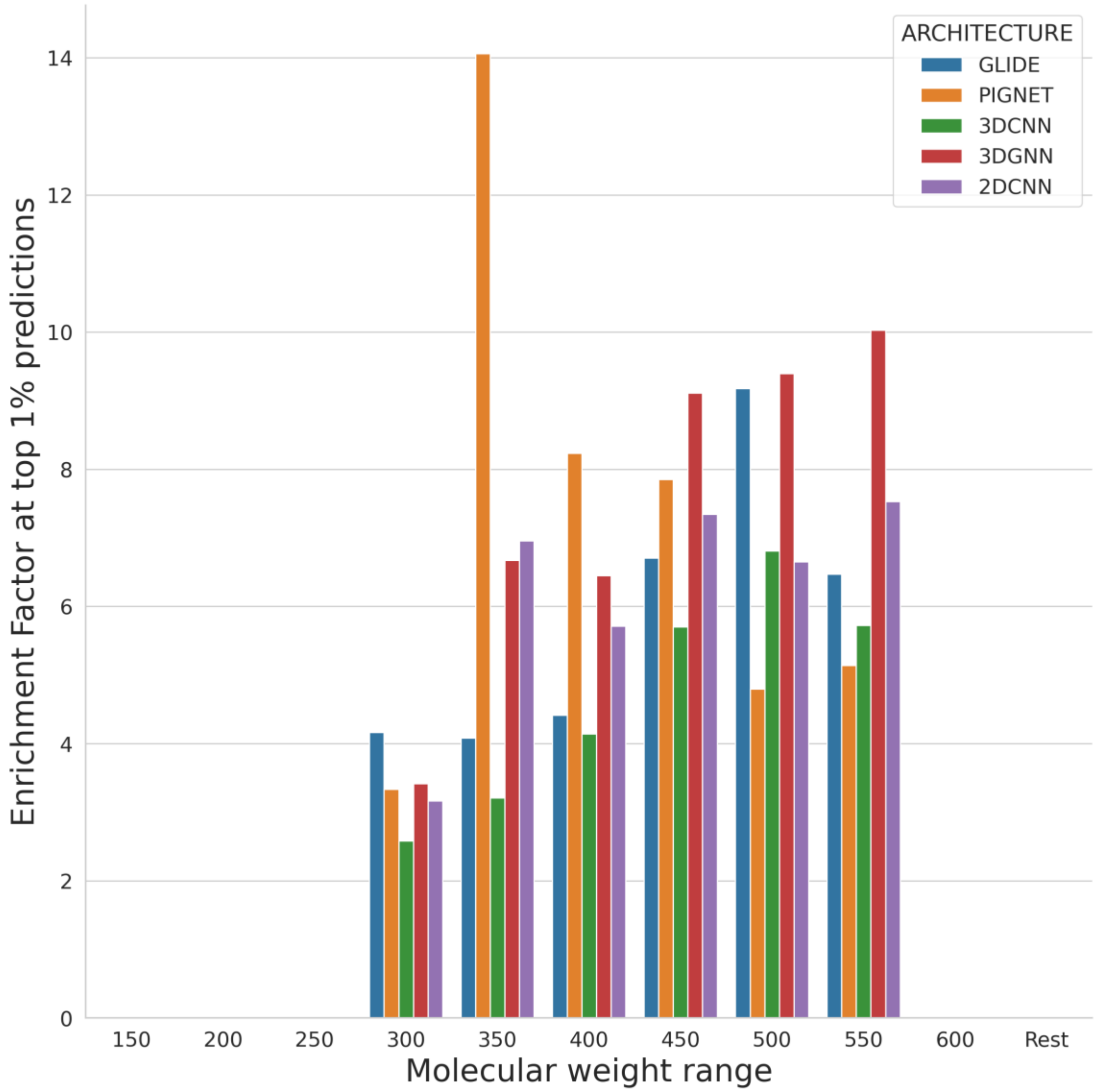

Figure 8. Active compounds enrichment factor in the top 1% predictions for various MW ranges using different architectures on the external ChEMBL test sets. The MW ranges are divided into bins separated by 50 Daltons, and each tick on the x-axis represents the range of its value minus 50.

MW bias. Hit enrichment was also analyzed for ChEMBL datasets. We stratified the data by MW and computed the average enrichment factor among datasets. As observed with internal LO test sets, our models (3DGNN, 2DCNN) showed consistently better enrichment factors compared to docking, especially the 3DGNN. In the range between 250-300 Daltons GLIDE showed better performance.

Hit diversity analysis. We also conducted the analysis of hit diversity for ChEMBL datasets. The average similarity is similar across the different architectures (Table 9, Figure 9). The diversity of hits in these datasets exhibits a pattern consistent with that observed in LO datasets. Diversity of hits appears to be more influenced by the characteristics of the dataset itself rather than by the specific architecture of the model.

Table 9. ChEMBL test datasets average similarity of top 100 predicted actives per architecture.

Architecture	**AVG_Similarity**
2DCNN	0.46
3DCNN	0.50
3DGNN	0.51
PIGNet	0.48
GLIDE	0.51

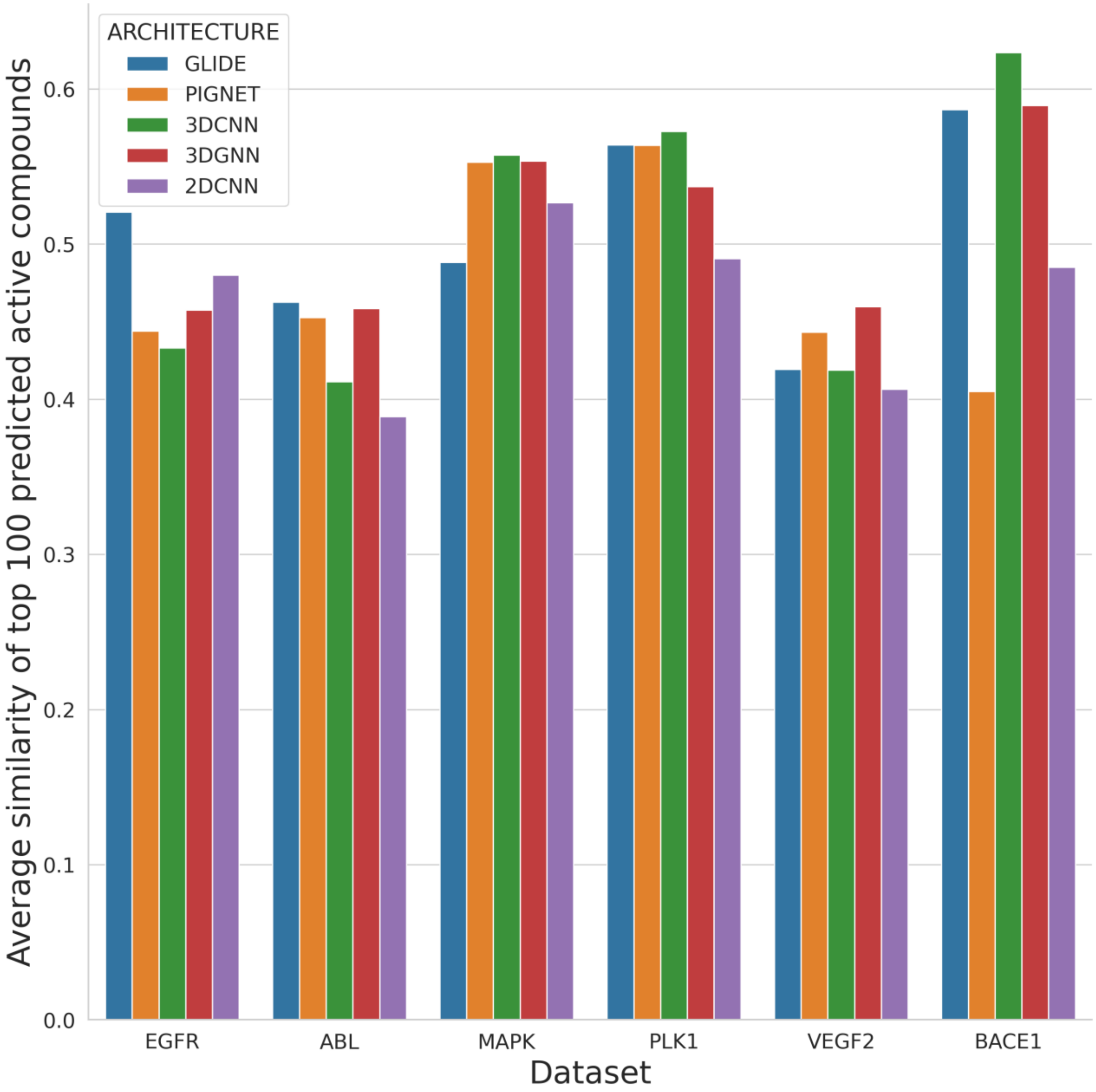

Figure 9. Average cross-similarity for the top predicted 100 active compounds per LO ChEMBL test set and model architecture.

DISCUSSION AND CONCLUSIONS

The prediction of absolute binding affinity using AI is a challenging task that has received considerable attention in recent years. Numerous efforts have been made to improve prediction performance, focusing on both the generation and implementation of new DL architectures and data improvement and augmentation. Despite significant progress, it is evident that the improvement of architectures has shown only a minor impact on performance, especially for external test sets. Our data showed the DL affinity prediction models do not perform better than docking with GLIDE on new external test sets (as measured by correlation coefficient). This is not evident from testing using the PDBbind core set that is commonly used in literature studies. We must consider truly new and challenging external datasets to reveal the issues. The limitations are likely centered on the available data, the underlying description or the overarching simplification of binding that ignores entropy, desolvation and other terms. These complex terms can be partly canceled if we were to consider bioactivity differences between similar compounds, a concept that motivated our previous work to learn based on differences in binding affinity[60] a topic recently revisited by others[61]. Regarding the data, the primary source of public data is the PDBbind database, which comprises curated protein-ligand complex structures available at the protein databank. However, it is worth noting that complexes formed by proteins and small molecules with reasonable quality are limited to approximately 5000 structures. This is a small number that does not adequately address the complex problem of predicting binding affinity from a data hungry perspective. Furthermore, biases that allow AI models to shortcut learning by correlating activity with simple data features like MW have been identified.

One of the most significant data limitations is the sparsity of the protein-ligand matrix, which presents a good representation of the protein space but limited chemical space. The chemical space of the ligands co-crystallized with a protein is usually diverse, which hinders the learning of the impact of small modifications in ligand structure and more importantly facilitates model overfitting on ligand chemical properties. As a result, DL models trained on PDBbind data often show poor performance when applied to external sets, especially drug discovery LO datasets, as studied here, where the data density is reversed with thousands of ligands (actives) for a limited number of proteins. Moreover, our results show how realistic measures of model performance can differ substantially versus testing with PDBbind core set, whose composition provides overoptimistic results.

To reduce the effect of data biases, data augmentation using inactive compounds with similar physicochemical properties have been included in training. Furthermore, we included pose decoys where active molecules in incorrect bound complexes were used for training. While both of these efforts are expected to improve model generalization by forcing the model to focus more on the interactions or learn that the features of the ligand alone are not responsible for the activity, it may be introducing noise via a ‘virtual binding’ mode and an assigned low pX_{50} (normalized between 0 and 3 in this work). From the DL architecture perspective, recent efforts make the models more generalizable by limiting learning to just interaction information, thus avoiding overfitting on the ligand, the protein or both. For example, Volkov et al. recently analyzed the problems related to the PDBbind dataset and DL models, concluding that for prospective applications, a model trained using interaction data might be more generalizable. However, the analyses performed in their work showed that even after subsampling the data to avoid overfitting, and considering only interactions in the graph representation, hidden biases were still present. There have also been efforts to incorporate more inductive biases into the models. For example, PIGNet, used in this work, is a physics-informed 3DGNN model in which the binding affinity is predicted as a sum of parameterized physics equations fed by interactions, corrected by a rotor term. The PIGNet models used in this work showed inconsistent results across datasets compared to other architectures. We think that the reason for this could be related to the difficulty for a few parameterized equations to learn all the aspects of binding affinity, which includes chemistry and the dynamical features like conformational sampling and solvation. Another interesting model is Dynaformer[62], a 3DGNN model trained with molecular dynamics data. Although the model accomplishes remarkable performance on the PDBbind coreset, performance on external sets is not provided, leaving doubts as to whether the model is learning physics thanks to the underlying simulations, or potentially learning new biases.

We have addressed the data sparsity problem by augmenting the data using drug discovery LO datasets and a rigorous docking protocol to ensure that resulting complexes are as close as possible to their hypothetical crystal structures. Interestingly, we unexpectedly observed an increase in PDBbind core set testing performance when training the model with drug discovery project LO data in comparison to training using only PDBbind data. More importantly we observed modest correlation for some of the external LO datasets where models trained with only PDBbind or GLIDE showed no trend. Interestingly, for the internal validation test sets the models performed rather poorly and similar in all cases but for the LO external test sets performance was also poor in terms of correlation of predicted and experimental activity but there was an improvement when using internal data. It must also be noted the increased difficulty of obtaining correlation for LO datasets where the distribution of activities across the activity range is imbalanced with most activity points concentrated in 2 to 3 pIC_{50} units. Considering the correlation between the predicted and experimental activity the overall performance was unclear if the DL models were outperforming docking. It is notable that docking (using GLIDE here) often performs well in the respective tests performed throughout our study. This contrasts with some reports where DL models often substantially outperform docking. However, upon close inspection, some of those studies use docking inappropriately, for instance to find a site in addition to the binding mode. As mentioned in the introduction, site identification was not the purpose for which docking algorithms were developed. Here, the docking performance is likely tending towards a better or more typical use case scenario because we are filtering to only retain the best scored, most reliable poses that may also conserve an MCS or certain interactions.

Overall, there were signs that for a virtual screening or hit finding application, the 3DGNN model trained including drug discovery LO training datasets achieved the highest performance with an EF at top 1% of 9.0 followed by the 2DCNN model and GLIDE with 7.3 and 6.6, respectively. The best model trained without LO datasets (3DGNN) achieved an EF at top 1% of 6.3. The 3DGNN was not only performing the best on average in terms of enrichment but also was more frequently the best model for a given target dataset. The models were resilient for challenging systems where there are ions and loop flexibility (PDE2, PDE10 and KDR). Hit enrichment is important, but also the quality/diversity of the hits. Molecular weight bias is also present in LO datasets, where hits are grown to improve their affinity during lead optimization. We evaluated the hit enrichment factors at different MW ranges and the similarity among the top scored active compounds. DL models were able not only to retrieve more hits, but also this was accomplished in all MW ranges. This suggests that the models can provide better enrichment at more attractive lower molecular weight ranges. Additionally, we measured the average similarity for top predicted active compounds. All methods show a similar average similarity demonstrating there is no clear advantage for any method in retrieving more diverse hits.

We also evaluated the model performance on ChEMBL datasets that were compiled using the same pipeline utilized for the drug discovery project LO datasets. Although the best model was the 3DGNN trained with internal data, the addition of the LO data did not offer an improvement as substantial as seen for the internal LO test sets. This finding can be explained by two main reasons: data quality and differences in chemical space. The data quality issue is mainly due to the difficulty in finding congeneric series of compounds with a representative analogue in a public domain crystal structure. This challenge compelled us to use constrained docking, which can result in inaccuracies in the computed best pose. Moreover, the advantage of training with the LO data may not be as relevant for the ChEMBL data as for the LO datasets due to increased difference in the chemical space between the two data sources.

We demonstrated that incorporating LO data during training can be beneficial in improving the performance of DL models for predicting binding affinity on datasets beyond the training domain. The increased data density facilitates the learning of SAR and reduces the biases associated to sparse data, like MW bias and the classification of activities by targets. Specifically, the Graph Neural Network (3DGNN) consistently showed better enrichment and MW diversity across most of the datasets in our external tests using LO project data compared to molecular docking scoring function. However, our models are not yet capable of providing a strong correlation between experimental and predicted activities. A particular limitation is the static representation of the system which underestimate the impact of the dynamics and the conformational changes, entropy and (de)solvation. The quantity and quality of data required to properly rank compounds within the small activity range are unknown, and it is unlikely that they will be available in the short or medium term. Instead, we believe that incorporating more physics into DL models and providing a complete picture of the thermodynamic cycle, either from simulations or other DL models, is a more promising direction to pursue.

ASSOCIATED CONTENT

Supporting Information. Supplementary tables 1-4.

AUTHOR INFORMATION

Corresponding Author

* Corresponding author: E-mail: gtresade@its.jnj.com. Tel: +32 1464 1569.

Author Contributions

The project was conceived by JCGT, MA, and GT. Experiments performed and analyzed by LC, JCGT, and GT. Manuscript written by JCGT, MA, and GT. Revision and approval of the manuscript by all authors.

ORCID

Gary Tresadern: 0000-0002-4801-1644

Jose Carlos Gomez Tamayo: 0000-0003-4709-6030

Mazen Ahmad: 0000-0003-4010-7227

ACKNOWLEDGMENTS

The authors acknowledge funding support from the European Commission for EU H2020 CompBioMed2 Centre of Excellence (grant no. 823712)

ABBREVIATIONS

ML, machine learning; MUE, mean unsigned error; PDB, protein data bank; RMSD; root mean square deviation; RMSE, root mean square error; SAR, structure activity relationship.

Supplementary information

Supplementary Table 1: Number of samples classified by their type used to train the different models in this work. Actives correspond to the sum of the data samples with affinity data measured. 2DCNN versions use both PDBBind refined and non-refined subsets (9395 samples). The remaining models used only the refined version of PDBbind. Models trained with LO data contain additional 16318 active data samples. Inactive compound complexes contain IBD molecules (831885 samples) plus cross-docked PDBbind samples (527682 samples). Pose decoy samples contain a decoy pose for each active compound complex for the 2DCNN models while the rest of the models served 292518 samples derived from PDBbind dataset.

Supplementary Table 2. Per-dataset disaggregated performance statistics in training test set splits.

Supplementary Table 3. Per-dataset disaggregated performance statistics for LO external test sets.

Supplementary Table 4. Per-dataset disaggregated performance statistics for chEMBL LO external test sets

References

1. Stanzione, F., Giangreco, I. & Cole, J. C. *Use of molecular docking computational tools in drug discovery*. *Progress in Medicinal Chemistry* vol. 60 (Elsevier B.V., 2021).

2. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. 1–29 (2022).

3. 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).

4. Diáz, L. *et al.* Monte Carlo simulations using PELE to identify a protein–protein inhibitor binding site and pose. *RSC Adv* **10**, 7058 (2020).

5. Wang, L. *et al.* Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc* **137**, 2695–2703 (2015).

6. Gapsys, V. *et al.* Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem Sci* **11**, 1140–1152 (2020).

7. Khalak, Y. *et al.* Alchemical absolute protein-ligand binding free energies for drug design. *Chem Sci* **12**, 13958–13971 (2021).

8. 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).

9. Gapsys, V. *et al.* Pre-Exascale Computing of Protein-Ligand Binding Free Energies with Open Source Software for Drug Design. *J Chem Inf Model* **62**, 1172–1177 (2022).

10. Khalak, Y., Tresadern, G., Hahn, D. F., De Groot, B. L. & Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *J Chem Theory Comput* **18**, 6259–6270 (2022).

11. Moore, J. H. *et al.* Automated relative binding free energy calculations: from SMILES to ΔΔG. *ArXiv* (2022).

12. Meli, R., Morris, G. M. & Biggin, P. C. Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. *Frontiers in Bioinformatics* **2**, 57 (2022).

13. Dhakal, A., McKay, C., Tanner, J. J. & Cheng, J. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Brief Bioinform* **23**, 1–23 (2022).

14. Jiménez, J., Škalič, M., Martínez-Rosell, G. & De Fabritiis, G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J Chem Inf Model* **58**, 287–296 (2018).

15. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34**, 3666–3674 (2018).

16. Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M. & Siedlecki, P. Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* **35**, 1334–1341 (2019).

17. Francoeur, P. G. *et al.* Three-dimensional convolutional neural networks and a crossdocked data set for structure-based drug design. *J Chem Inf Model* **60**, 4200–4215 (2020).

18. Wang, Z. *et al.* OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity Based on Residue-Atom Contacting Shells. *Front Chem* **9**, 913 (2021).

19. Stafford, K. A., Anderson, B. M., Sorenson, J. & Van Den Bedem, H. AtomNet PoseRanker: Enriching Ligand Pose Quality for Dynamic Proteins in Virtual High-Throughput Screens. *J Chem Inf Model* **62**, 1178–1189 (2022).

20. Meli, R., Anighoro, A., Bodkin, M. J., Morris, G. M. & Biggin, P. C. Learning protein-ligand binding affinity with atomic environment vectors. *J Cheminform* **13**, 1–19 (2021).

21. Moon, S., Zhung, W., Yang, S., Lim, J. & Kim, W. Y. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chem Sci* **13**, 3661–3673 (2022).

22. Shen, C. *et al.* Boosting Protein-Ligand Binding Pose Prediction and Virtual Screening Based on Residue-Atom Distance Likelihood Potential and Graph Transformer. *J Med Chem* **65**, 10691–10706 (2022).

23. 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).

24. Wang, D. D., Chan, M. T. & Yan, H. Structure-based protein–ligand interaction fingerprints for binding affinity prediction. *Comput Struct Biotechnol J* **19**, 6291–6300 (2021).

25. Liu, Q. *et al.* OctSurf: Efficient hierarchical voxel-based molecular surface representation for protein-ligand affinity prediction. *J Mol Graph Model* **105**, (2021).

26. Feinberg, E. N. *et al.* PotentialNet for Molecular Property Prediction. *ACS Cent Sci* **4**, 1520–1530 (2018).

27. Lim, J. *et al.* Predicting Drug-Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *J Chem Inf Model* **59**, 3981–3988 (2019).

28. Torng, W. & Altman, R. B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J Chem Inf Model* **59**, (2019).

29. Karlov, D. S., Sosnin, S., Fedorov, M. V. & Popov, P. GraphDelta: MPNN Scoring Function for the Affinity Prediction of Protein-Ligand Complexes. *ACS Omega* **5**, 5150–5159 (2020).

30. 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).

31. Ahmed, A., Mam, B. & Sowdhamini, R. DEELIG: A Deep Learning Approach to Predict Protein-Ligand Binding Affinity. *Bioinform Biol Insights* **15**, (2021).

32. Jones, D. *et al.* Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *J Chem Inf Model* **61**, 1583–1592 (2021).

33. Liu, Z. *et al.* PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).

34. Chen, L. *et al.* Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One* **14**, e0220113 (2019).

35. 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).

36. Sieg, J., Flachsenberg, F. & Rarey, M. In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. *J Chem Inf Model* **59**, 947–961 (2019).

37. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).

38. Yang, Z., Zhong, W., Zhao, L. & Yu-Chian Chen, C. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chem Sci* **13**, 816–833 (2022).

39. Pérez-Benito, L., Casajuana-Martin, N., Jiménez-Rosés, M., Van Vlijmen, H. & Tresadern, G. Predicting Activity Cliffs with Free-Energy Perturbation. *J Chem Theory Comput* **15**, 1884–1895 (2019).

40. Furtmann, N., Hu, Y., Gütschow, M. & Bajorath, J. Identification and analysis of the currently available high-confidence three-dimensional activity cliffs. *RSC Adv* **5**, 43660–43668 (2015).

41. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **47**, D930–D940 (2019).

42. Li, Y., Han, L., Liu, Z. & Wang, R. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *J Chem Inf Model* **54**, 1717–1736 (2014).

43. Sánchez-Cruz, N., Medina-Franco, J. L., Mestres, J. & Barril, X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* 1–7 (2020) doi:10.1093/bioinformatics/btaa982.

44. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

45. Friesner, R. A. *et al.* Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **49**, 6177–6196 (2006).

46. Scantlebury, J., Brown, N., Von Delft, F. & Deane, C. M. Data Set Augmentation Allows Deep Learning-Based Virtual Screening to Better Generalize to Unseen Target Classes and Highlight Important Binding Interactions. *J Chem Inf Model* **60**, 3722–3730 (2020).

47. Shen, C. *et al.* The impact of cross-docked poses on performance of machine learning classifier for protein–ligand binding pose prediction. *J Cheminform* **13**, 1–18 (2021).

48. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **55**, 6582–6594 (2012).

49. Zhang, X. *et al.* Topology-Based and Conformation-Based Decoys Database: An Unbiased Online Database for Training and Benchmarking Machine-Learning Scoring Functions. *J Med Chem* **66**, 9174–9183 (2023).

50. Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* **27**, 221–234 (2013).

51. Jacobson, M. P. *et al.* A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**, 351–367 (2004).

52. Bas, D. C., Rogers, D. M. & Jensen, J. H. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins* **73**, 765–783 (2008).

53. Greenwood, J. R., Calkins, D., Sullivan, A. P. & Shelley, J. C. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J Comput Aided Mol Des* **24**, 591–604 (2010).

54. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model* **53**, 1893–1904 (2013).

55. Landrum, G. RDKit Documentation. *Read Writ* (2011).

56. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J Chem Inf Model* **50**, 742–754 (2010).

57. Manimegalai, P. *et al.* 3D Convolutional Neural Network Framework with Deep Learning for Nuclear Medicine. *Scanning* **2022**, (2022).

58. Wu, Z. *et al.* A Comprehensive Survey on Graph Neural Networks. *IEEE Trans Neural Netw Learn Syst* **32**, 4–24 (2021).

59. Reiser, P. *et al.* Graph neural networks for materials science and chemistry. *Commun Mater* **3**, (2022).

60. Jiménez-Luna, J. *et al.* DeltaDelta neural networks for lead optimization of small molecule potency. *Chem Sci* **10**, 10911–10918 (2019).

61. McNutt, A. T. & Koes, D. R. Improving ΔΔG Predictions with a Multitask Convolutional Siamese Network. *J Chem Inf Model* **62**, 1819–1829 (2022).

62. Min, Y. *et al.* From Static to Dynamic Structures: Improving Binding Affinity Prediction with a Graph-Based Deep Learning Model. (2022).