Assessment of molecular dynamics time series descriptors in protein-ligand affinity prediction.

Jakub Poziemski [1], Artur Yurkevych [2], Paweł Siedlecki [1]

[1] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
[2] Institute of Chemistry, University of Silesia in Katowice, Katowice, Poland

Abstract

The advancement of computational methods in drug discovery, particularly through the use of machine learning (ML) and deep learning (DL), has significantly enhanced the precision of binding affinity predictions. Despite progress in computer-aided drug discovery (CADD) accurate prediction of binding affinity remains a challenge due to the complex, non-linear character of molecular interactions. Generalizability continues to limit these models, with performance discrepancies noted between training datasets and external test conditions. This study explores the integration of molecular dynamics (MD) simulations with ML to assess its predictive performance and limitations. In particular MD simulations offer a dynamic perspective by depicting the temporal interactions within protein-ligand complexes, potentially bringing additional information for affinity and specificity estimates. By generating and analyzing over 800 unique protein-ligand MD simulations, we evaluate the utility of MD-derived descriptors based on time series in enhancing predictive accuracies. The findings suggest specific and generalizable features derived from MD data and propose approaches to augment the current *in silico* affinity prediction methods.

Introduction

The quest for efficient drug discovery is critically dependent on the precision of binding affinity predictions, an area where computational methods have both flourished and found limitations [1]. Computer-aided drug discovery (CADD) techniques have made an impact on the pharmaceutical industry by enhancing the efficiency of the drug development process, reducing time, cost, and labor. Despite these advancements, accurate prediction of binding affinity continues to pose a considerable challenge, often bottlenecked by the inherent complexities of molecular interactions [2,3].

Rapid and continuous progress in machine learning (ML) and deep learning (DL) has shown promise in overcoming some of these hurdles [4–6]. They have improved our ability to interpret and utilize large datasets, revealing intricate non-linear patterns and relationships that are crucial for the stability of protein-ligand complexes. Current state-of-the-art methods achieve Pearson correlation coefficient (R_p) around 0.7-0.85 in the CASF2016 benchmark [7,8], which is a significant improvement compared to the classical scoring functions used previously. Despite this achievement, challenges remain, particularly with the generalizability of these models, which often perform well on training datasets but show diminished accuracy on external test sets or in individual studies. There are many hints of why this plateau has occurred and is difficult to overcome. One important consideration is, while useful, traditional static computational approaches like molecular docking only provide a limited view by capturing snapshots of molecular complexes without their temporal dynamics [9,10].

Molecular dynamics (MD) simulations, although not without their own hurdles, introduce a vital temporal dimension to protein-ligand complex studies, offering a dynamic perspective that is more reflective of the actual biological processes. Such simulations allow for more detailed observations of how drug molecules interact with biological targets over time, which can be essential for understanding both the affinity and specificity of interactions [11–13]. This dynamic insight helps in recognizing conformational changes during binding and unbinding events, which are essential in many applications of the drug design process.

The integration of MD simulations with ML and DL has been hinted as an attractive path to achieve an improved level of predictive performance. Over the last years it has been tested and applied with varying success in different drug discovery tasks and specific campaigns. Riniker [14] developed and applied Molecular Dynamics Fingerprints (MDFP) to the problem of solvation free energies prediction. The ML models were trained with distribution properties of potential-energy components, radius of gyration (Rg), and solvent-accessible surface area (SASA) extracted from 5ns MD simulations performed on 426 small molecules derived from the FreeSolv database [15], for the prediction of solvation free energies as well as different partition coefficients. MDFP showed predictive performance useful for computer-aided lead optimization and analogue prioritization. In the case of affinity prediction, Ash and Fourches in 2017 [16] analyzed 87 ERK2- docked ligand complexes by computing chemical descriptors derived from 20ns molecular dynamics (MD) trajectories. They showed that models trained on MD derived descriptors were able to distinguish the most active ERK2 inhibitors from the moderate/weak actives and inactives. They claimed that the descriptors extracted from MD trajectories are highly informative and, having little correlation with classical 2D/3D descriptors, could augment chemical libraries screening tasks, candidate design and lead prioritization. A similar concussion was presented by [17], who performed molecular docking of 43 compounds associated with caspase-8, with consecutive 10-ns MD simulations of top scoring complex for each ligand. They investigated 770 2D and 115 3D

descriptors together with 4 descriptors extracted from MD simulations: solvent accessible surface area (SASA), radius of gyration (Rg), potential energy and total energy, in the form of mean and standard deviation (8 descriptors in total). They reported that ML models trained on MD data had the most balanced accuracies and AUC values, compared to the 2D and 3D descriptor models, and that models using a combination of 3D and MD descriptors had the best performance. A counter experiment was performed by [18] using the BCR-ABL tyrosine-kinase and 15ns MD simulations of Imatinib and a large series of its derivatives. In conclusion they stated that incorporating time-series-based MD matrices could not improve the binding affinity prediction ability of the DNN and random forest (RF) QSAR models. However, their models did have reduced prediction error, indicating that the MD trajectories contain both useful information and noises, with the negative effect from the noises becoming stronger as the number of snapshots increases. An approach to compare different ML models trained on descriptors obtained from MD trajectories was presented by [10]. Using three different targets and a maximum of 433 complexes predicted by docking per target, the results for MD augmented approaches were greatly dependent by target. The paper concludes the use of MD does not generally improve screening results and may only be justified in certain cases. Given that the models for MD data were trained for descriptors generated from each frame, this may have been a challenge for simple ML models due to the small amount of data. In addition, the low MM/PBSA and Glide scores suggest that the analyzed collections were rather difficult.

Current research indicates the complexity of leveraging molecular dynamics (MD) data, suggesting it is target specific and can depend on the noise to signal ratio, e.g. number of frames, the length of MD simulation, etc. It is difficult however to draw definite conclusions as only a handful of targets have been tested so far. Given the structural diversity of the ligand-protein complexes, it is far too few to assess the overall usefulness of MD simulations in large scale prediction of activity. Can the non-linear features of P-L complexes extracted from MD data be used to enhance the predictive accuracy of binding affinity prediction? Can it lead to the identification of novel features useful for such prediction?

To answer some of these questions, in this study we have generated the largest set of MD simulations to date, encompassing a broad array of protein-ligand complexes. Constructing a large representative set of MD simulations is challenging due to the high computational cost and difficulty associated with molecular system preparations. In addition, given the small number of training examples (complexes suitable for conducting simulations) and a large amount of MD data from each simulation, careful filtering and feature selection is of utmost importance. Therefore we generated a comprehensive set of descriptors that utilize various aspects of MD-derived data, and implemented a rigorous feature selection mechanism to tailor the number of features to the analytical methods employed. By training ML models on MD-derived data from over 800 unique protein-ligand complexes, we seek answers to the feasibility of greater accuracy. In this work we ask the following questions: 1) What are the complex specific and simulation specific features influencing models' affinity prediction outcomes? 2) Whether MD-derived simulation data is beneficial and how it generalizes on a large scale? 3) Can the MD-derived descriptors augment and/or replace current crystallographic derived descriptors? Based on our findings we present general guidelines and suggestions on how to augment the *in silico* affinity prediction pipelines in the context of MD simulations treated as time series.

Materials and methods

Dataset compilation

The study utilizes protein-ligand crystallographic complexes sourced from the PDBBind collection v2020 [19]. The effort was to compile a diverse dataset with the goal to minimize various ligand- and target based biases. With this in mind, the selection criteria considered only complexes with well-defined active sites paired with ligands with unambiguous, experimentally determined affinity values (in the form of negative log Ki, Kd and IC50 values). Criteria for exclusion included peptide ligands, multi-chain, protein complexes, trans-membrane proteins, and any proteins where a metal ion is present in the active site (defined as 6A away from any ligand atom). To maintain diversity, a cap was set at 18 protein complexes per target, identified by a common UniProt identifier. All complexes were required to have all amino acids crystallized and identified.

Ligand diversity

There are 862 complexes in the MDD dataset, around half of them with a single ligand, 71 targets with 2 to 5 ligands, 31 with 6-10 and 22 with than 10 (Figure 1, A). The affinity range of ligands in the MDD dataset is described roughly by a normal distribution (Figure 1, B) when the logarithm scale is used (i.e. pKi/pKd/pIC50). If we consider 1uM affinity (6 on the logarithmic scale) as a threshold for defining active/inactive classes, the MDD dataset shows an almost equal distribution of active and inactive compounds (420 ligands below 6, and 444 equal or above 6). From the physicochemical point of view 96% of MDD ligands comply with RO5 (Figure 1, C-H). The structural diversity of the MDD ligands was measured with the ECFP4 (1024 bits) fingerprint. Considering all ligands in the dataset, the mean Tanimoto distance between them is Tc=0.11, showing a low overall structural similarity of the whole small molecule space. Similarity of ligands within their targets is also low, the mean Tc distance is 0.30.

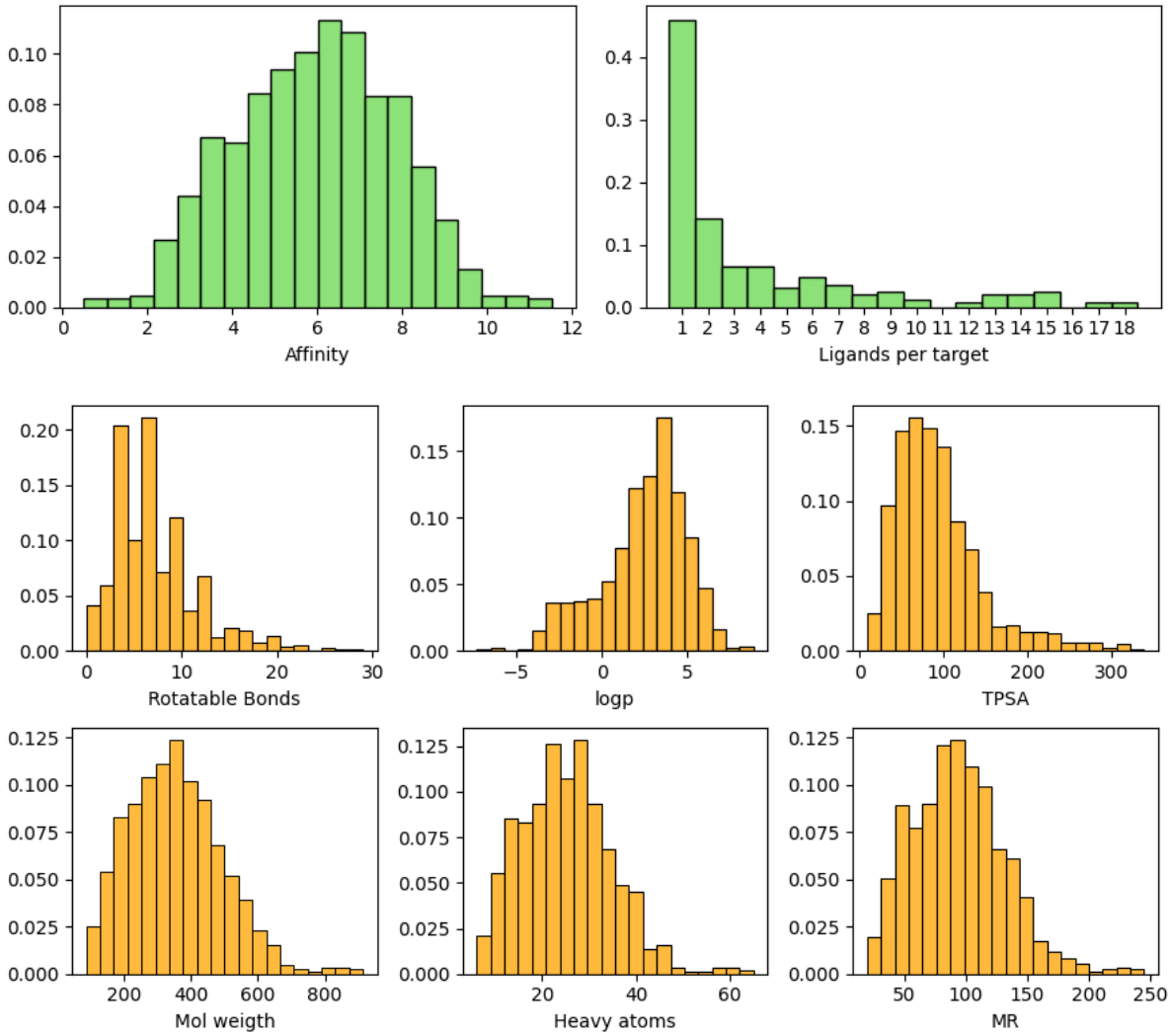

Figure 1: Ligand physicochemical features distribution in the MDD dataset. Selected ligand descriptor distributions are compared to the protein-ligand part of the PDBBind dataset. Ligand per target depicts the fraction of targets with a given number ligand complexes.

Target diversity

The MDD dataset comprises 231 targets, compared both from a functional and structural perspective. Roughly around ⅔ of the MDD targets are enzymes with an assigned EC number, with hydrolases and transferases composing over ½ of the MDD. Around ⅓ of the MDD targets are non-enzymatic proteins, with the largest group described as “transport proteins” by GO Biological Process keywords. More details of the functional composition of MDD are available in Supplementary materials file: MDD_targets.ods

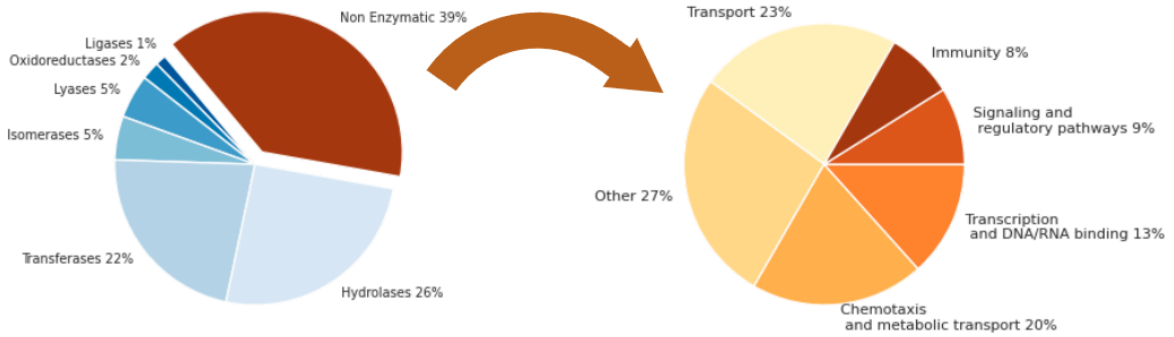

Figure 2: Functional characterization of MDD targets. The “Non-enzymatic” part of MDD (⅓ of the targets) are described on the right chart with 5 distinct biological processes (GO annotation). Around 10% of all MDD targets are non-enzymatic proteins with other functions.

Binding site similarity was assessed with DeeplyTough [20], and presented on Figure 2. We chose this method as it combines both structural similarity and ligand preference similarity into a single comparison value which, after normalization, is easy to interpret and process. The median similarity score between all MDD targets is equal to 0.62 (after normalization), showing rather low binding pocket similarity (Figure 3, A). In detail, around 8.5% of the compared structural pairs are highly similar (comparison value greater >= 0.95) and 23% of pairs are dissimilar (values < 0.5). Target binding sites were also compared with respect to hydrophobic residues, size (area), and mobility (RMSF) (Figure 3, B-D). Overall the MMD targets show a good balance of the above features with close to normal distributions. We note the RMSF values are mostly between 0,5-1,5A, suggesting the MDD targets do not experience major conformational changes, at least during 200ns MD simulations with their ligands.

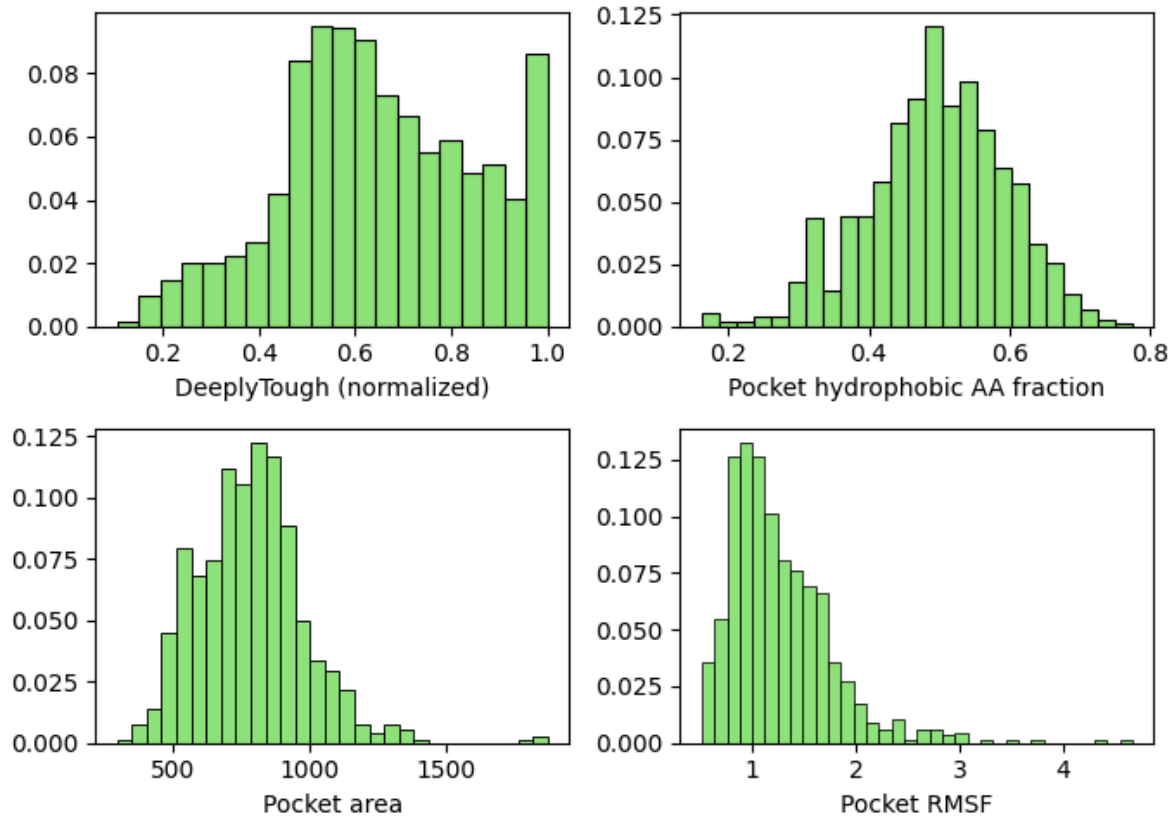

Figure 3: Comparison of binding site properties of MDD targets. Comparison with respect to structural similarity and ligand preference similarity (DeeplyTough), hydrophobic residues, size (area), and mobility (RMSF).

MD simulation procedure

All protein-ligand complexes selected for molecular dynamics simulations were prepared following a standardized protocol. Missing atoms in the protein structures were added using the PDBFixer tool [21]. Protein targets were parameterized using the AMBER99SB-ILDN force field, while ligand parameterization was conducted with the ANTECHAMBER module within the ACPYPE tool [22]. For the ligands, partial charges were derived to match the quantum-mechanically generated electrostatic potential via the Restrained Electrostatic Potential (RESP) method [23], and the remaining parameters were aligned using the GAFF2 force field. The aim of this procedure was to provide a generic method for complex parameterization applicable to a wide variety of protein-ligand complexes.

Molecular dynamics simulations were executed using the GROMACS [24]. The simulations were set up in a cubic simulation box with periodic boundary conditions, employing a TIP3P water model within an electrostatically neutral environment. The simulation protocol included an initial minimization cycle, followed by temperature equilibration in the NVT ensemble and pressure equilibration in the NPT ensemble. Production simulations were conducted over a 200 ns timeframe, with a timestep of 100 ps.

Representation

Given the relatively small data set, we decided to use simple ML methods. In addition, for the same reason, we tried to limit the number of features used in the learning process as much as possible. Another important aspect in the selection of the input set of features was the choice of features that would describe the target-ligand complex as from different perspectives, hence we used descriptive types features: physicochemical properties, geometries and compositions of the ligands, as well as the binding pocket together with features describing the target-ligand interactions. In the case of molecular dynamics data we calculated features describing the movement of the active site and the ligand, at the same time bearing in mind to use as few features as possible. Taken together we calculated 29 ligand descriptors (calculated with [25,26], 23 pocket descriptors (with [25,27], 13 interaction descriptors (calculated with PyLIF [28]) and 8 motion descriptors (derived from [27]). Details of the specific types and calculations are attached in the Supplementary Materials section "List of descriptors".

In this work we use two types of structural data from the same experimentally determined structures: the "crystallographic poses" (single frame data), and the 200ns MD simulation frames (saved every 100ps - multiframe data). For the multiframe data we can use `tsfresh` [29] to calculate time series features for each descriptor. A multistep procedure was executed:

1. For each decryptor (which can be represented in a time series), we calculate all 788 time series descriptors (ts_descriptors) supported by `tsfresh` package (v. 0.20.1).
2. For each ts_descriptor, its p-value is determined in terms of statistical significance against experimentally determined affinity (source from PDBBind), using a univariate test, with FDR set to 0.001 (see Figure 4).
3. Next, a trimming procedure is used. For each of the 75 features types (see: tsfresh features), the ts_descriptor with the lowest p-value is selected. Some ts_descriptors are parametric in nature, using different thresholds one can generate several versions of the same ts_descriptor. For such a group, also the one with the lowest p-value was selected. At this stage, there could be a maximum of 75 ts_descriptors per complex descriptor.
4. To avoid caveats in training, such as correlation between predictive features, correlation trimming was applied. All descriptors and ts_descriptors were tested for correlations with each other. Correlated descriptors were dropped if PCC was >= 0.8. Among a group of correlated ts_descriptors, the one with the highest cardinality was left.
5. In the final filtering step, for each descriptor, the ts_descriptor with the highest mutual information score (MI SC) between itself and the experimentally determined affinity value was selected. As a result, each descriptor is described with at most one ts_descriptor (Figure 4, bottom left).

In total, for the crystallographic set we calculated a maximum of 63 descriptors (ligand, target, and interaction, some might be removed due to cross-correlation, see point 4, above) together with the ECFP4 fingerprint (1024bits). For the MD set we calculated the same 63 descriptors, 7 motion descriptors and up to 51 ts_descriptors.

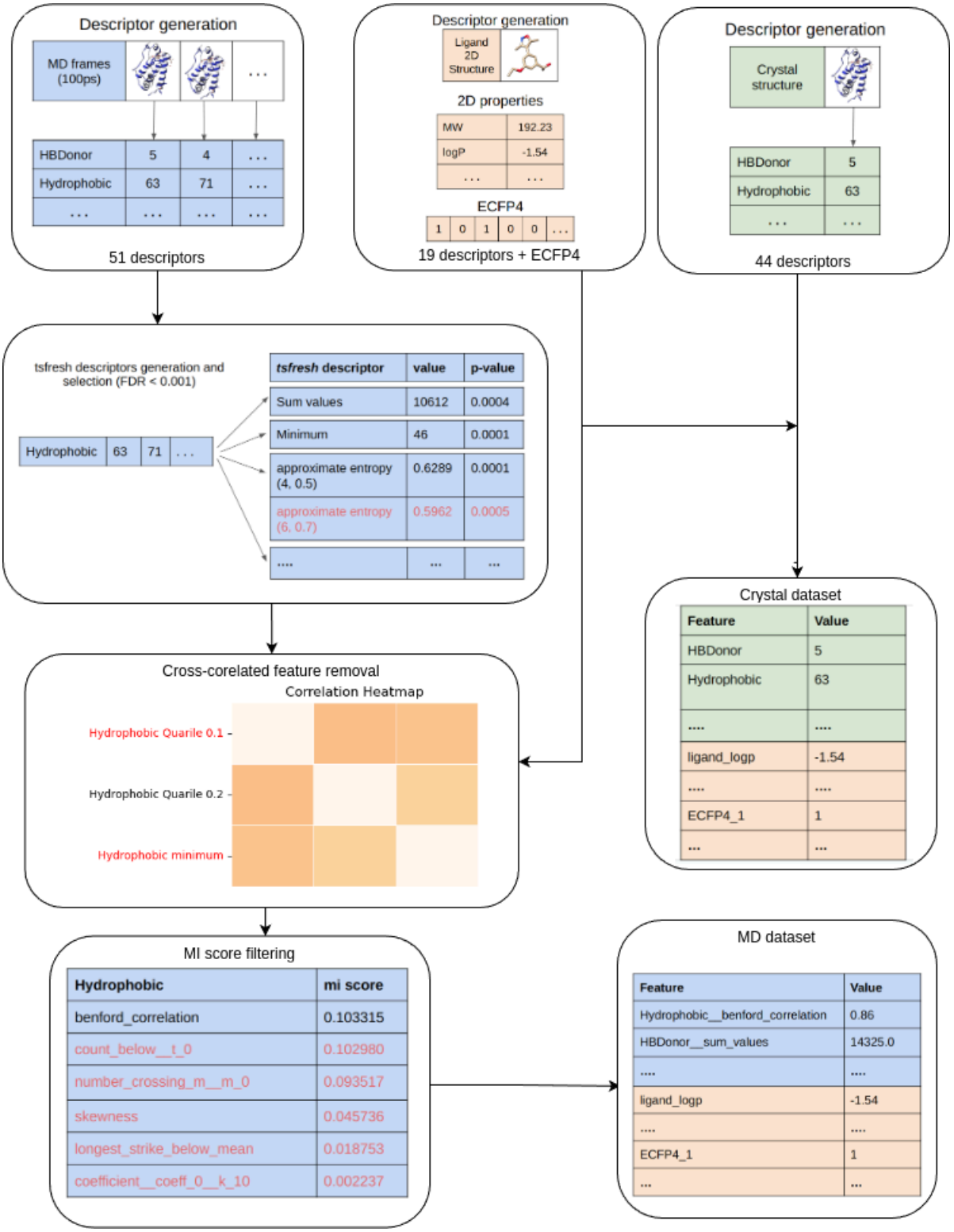

Figure 4. Flowchart of descriptors and ts_descriptors selection procedure. Color code: light green - crystallographic set descriptors, light blue - multiframe MD set descriptors and ts_descriptors, light orange - ligand descriptors. See section “Representation” for more details.

Data splits, training and testing

Two different ways of splitting the target data were used: random split and target split. In the random split, complexes were randomly allocated to the training and test sets, at a ratio of 4:1 (80% of complexes to the training collection, 20% to the test collection). In the case of target split, UniprotIDs were utilized to split the complexes, in the same proportions as for the random split. The latter split is more difficult as there should not be any similar training and testing examples (i.e. no identical protein targets between train and test sets). The target split can therefore be used to approximate the generalization potential.
Scaffold split, implemented in DeepChem [30] was used to split ligand structures deposited in MDD. ScaffoldSplitter DeepChem package divides molecules into groups, based on their Bemis-Murcko scaffold, and puts the smallest groups into the test set. Such a division is fully deterministic; the rarest scaffolds always constitute the test set. This type of scaffold split is more challenging than random splits as it tests more thoroughly the generalizability to new or less abundant areas of chemical space.
Model parameters were selected using 5-fold cross-validation (CV) on the training set. Three different models were trained using the MDD dataset; Random Forest and SVM with [31] and XGBoost with [32]. The models were fitted to the training sets and evaluated on the test sets. The whole split and training procedure was repeated 20 times. Throughout this work the test set results are reported as boxplots, with mean (triangle) and median (horizontal line inside the boxplot) values.

Results

Dataset establishment

In order to test our hypothesis that MD descriptor-augmented machine learning models may improve affinity prediction tasks, a proper selection of training examples was needed. On the one hand one wants to compile a large dataset with diverse target-ligand complexes to avoid skewed feature distributions. Skewed distributions can have an impact both on training procedures, generalizability and may result in improper assessment of performance [33]. On the other hand short MD runs do not usually account for allosteric changes, substantial pose alterations, etc. Therefore a set of filters were applied to over 19K target-ligand structures deposited in the PDBBind database. Briefly, using the procedure described in the Material and Methods section, we selected medium size globular proteins complexed with ligands having unambiguous, experimentally determined affinity values (pKi, pKd, pIC50). We use PDBBind as the primary source of structures, affinity measurements, and binding site definition.

Next we introduce an automatic MD procedure (see Materials and Methods section “MD simulation” for details). In general we wanted our procedure to be as generic as possible, without the need of any structural alterations or expert knowledge. This would in theory assure such procedure is transferable and generalizable, and could be used for other related tasks such as SBVS, or de novo design. To decrease the chance of errors in the automatic MD procedure we excluded targets with missing AA, cofactors and ions in the binding site.

After the filtering step, the remaining complexes were prepared for conducting MD simulations. Targets were parameterized with AMBER99SB-ILDN force field [ref], while topology parameters for organic chemical compounds were generated with ACPYPE [22]. If there were errors in the ACPYPE procedure such a complex was generally rejected; in a few cases we were able to manually rescue the PDB structure by small fixes. Finally, we arrived at 876 complexes for which we could perform 200ns MD simulations and collect simulation data. We call this set the “Molecular Dynamics Dataset” or MDD for short.

Baseline performance

To assess the difficulty of the MDD, affinity prediction performance was calculated with selected models for which code and training procedures were available in public repositories (Table 1). To conduct a fair comparison with minimized data leakage events, all models were trained on the PDBBind dataset (v2016) with 862 MDD complexes excluded.

Data type and model	PCC	RMSE	features
OnionNet2	0.75	1.26	defined in [6]
PLEC-NN	0.72	1.40	defined in [34]
RF-Score v2	0.63	1.58	defined in [4]
RF-Score v1	0.59	1.67	defined in [4]
NN-Score	0.59	1.66	defined in [35]
Vina	0.49	-	defined in [36]

Table 1. Performance of selected affinity prediction methods on the MDD subset. Pearson correlation coefficient (model vs experimental affinity).

We compared the MDD results (Table 1) to published assessments done with CASF2016 (Table 2). Various methods with increasing complexity and time necessary for training are displayed in Table 1 and Table 2, from simple classical scoring functions such as Vina, through machine learning models, up to sophisticated deep- and graph neural networks. The MDD results show a consistent and rather equal drop in performance compared to published results, independent of the tested methods. The observed decrease in performance may be multifacet: CASF2016 is a relatively small dataset (216 complexes) so it might be easier to optimize for performance. Also, the MDD has around four times more compleses than CASF2016; excluding MDD targets might decrease the availability of information needed for high affinity prediction performance. Importantly, Table 2 shows the performance of a model based on descriptors selected with our procedure: “descriptor model (XGB)” (see Materials and Methods Representation section and Figure 1, orange and green elements). Its affinity prediction is on par with some of the best, highly sophisticated methods. The results highlight that a relatively simple ML model can show a similar level of performance compared to specialized neural networks.

Model name	Description	PCC	RMSE	Training size	References
OnionNet2	CNN trained on contact descriptors	0.86	1.16		[6]
TopBP	Topological Descriptors with GBT	0.86	1.19	3 767	[37]
SS-GNN	Graph Neural Network	0.85	1.18	15 394	[38]
Descriptor model (XGB)	XGBoost	**0.85**	**1.202**	**12 866**	
DCML	Dowker complex based machine learning	0.84	1.25	3 772	[39]
OPRC-GBT	Ollivier persistent Ricci curvature	0.84	1.25	3 772	[40]
PLANET	Graph Neural Network	0.82	1.24	15 616	[41]
K_{DEEP}	Convolution Neural Network	0.82	1.27	3 767	[42]
PLEC-NN	Extended Connectivity FP & Neural Network	0.82	1.25		[34]
OnionNet	Convolutional Neural Network	0.82	1.27	11 906	[43]
RF-Score v1	Random Forest	0.80	1.39	3 767	[4]
Pafnucy	Convolution Neural Network	0.78	1.42	11 906	[5]
Vina	Hybrid empirical scoring function	0.60	-	-	[44]

Table 2: Affinity prediction performance using models of increasing complexity, tested with CASF2016 benchmark. The use of simple models on crystallographic data can yield comparable results to the use of complex neural networks based models.

Encouraged by these results we sought to explore the relatively simple ML based models, trained both on descriptors extracted from crystallographic and from MD simulation experiments, to assess the potential benefit of such an approach in the task of large scale affinity prediction.

Simulation length vs model performance

To determine what is the optimal length of MD simulations for information extraction in our setup, we defined 5 timescales between 10ns up to 200ns, and trained 3 types of models: RF, XGB and SVM with increasing MD lengths. The results for all three models were very similar (data not shown), therefore for further work the RF model was chosen, due to its simplicity and low tunability. Figure 5 summarizes the results obtained from the described

above experiments. The results show varying correlations, both with random and target splits. The best overall performance for both types of splits is achieved for a 20ns simulation length, as measured with PCC, r2 and mse values. In the following sections, all analyses refer to the RF models trained on 20ns molecular dynamics, unless mentioned otherwise.

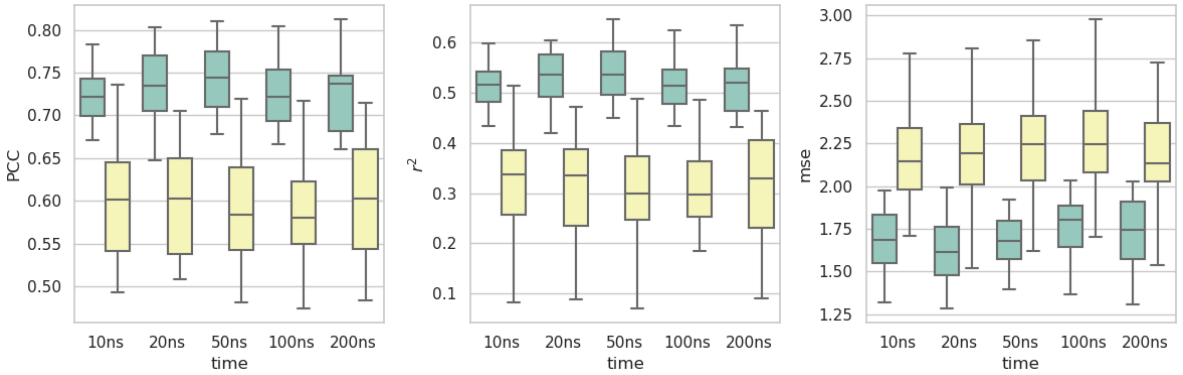

Figure 5: Dependence of MD simulation length on model performance. RF models trained with trajectories of different lengths (from 10ns to 200ns) tested on two types of data splits: random (green) and target (yellow). Increasing the length of the MD simulation does not improve the model performance results. 20ns trajectories show good overall performance for both random and target splits, therefore were used for subsequent training and analyses.

To assess which features could be responsible for the gain in performance of models trained on 20ns compared to 200ns we analyzed individual time series descriptors (ts_descriptors). The results show a significant gain in around 40% of tested ts_descriptors (14 out of 36, with $\Delta > 0.1$). Comparable correlations are registered for 12 ts_descriptors ($\Delta < 0.03$). Taken together our results suggest that longer MD runs may contain more noise vs information useful for the affinity prediction problem. In the assessed timescale, this may indicate that short MD simulations may be enough to capture useful steric conformation changes. Similar results have been present in works of others, concerning single targets [10,16,17].

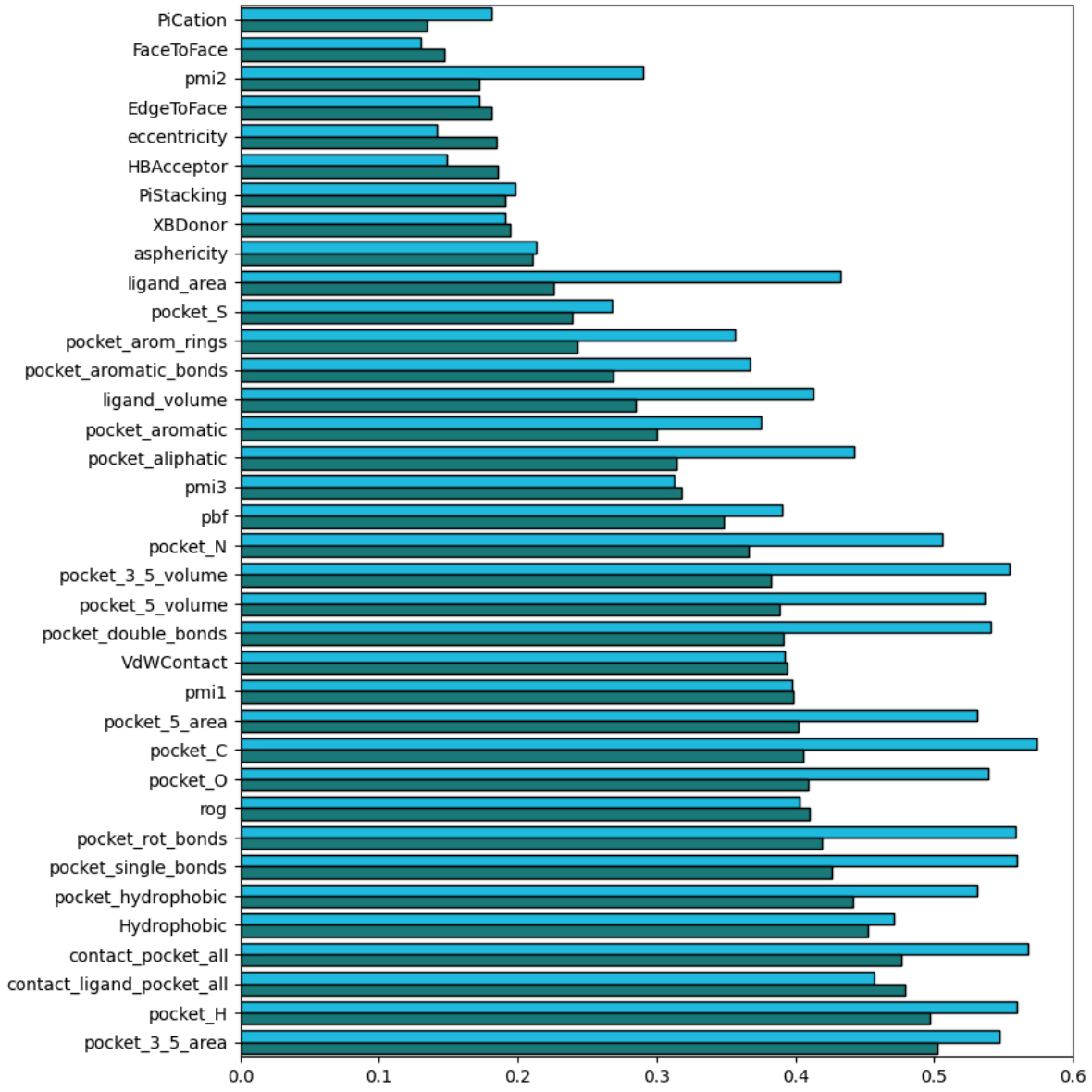

Figure 6: Correlations between affinity prediction and time series descriptors of the protein-ligand complexes. ts_descriptors derived from 20ns (light blue) and 200ns (dark blue) molecular dynamics simulations.

MD augmented representation

To understand the influence of time based descriptors on the model performance we tested two types of models with respect to affinity prediction tasks. In particular, we assessed the overall performance of a model trained with crystallographic descriptors only, and augmented with MD derived ts_descriptors, with respect to different target and ligand data splits. The results are shown on Figure 7. For randomly split data there is no clear difference in performance; both models achieve comparable results with respect to PCC, r2 and similar SD. However, in the case of a more challenging target split, an advantage of the model trained on dynamic descriptors can be seen, along with a smaller SD. A similar trend is even more visible with scaffold split, both with PCC and r2. Here the Bemis-Murcko scaffolds are used to split ligands into groups, next the test set is formed from the least abundant scaffold

groups. Such an approach is more challenging than random splits, with a more thorough test on generalizability, especially with respect to uncharted chemical space.

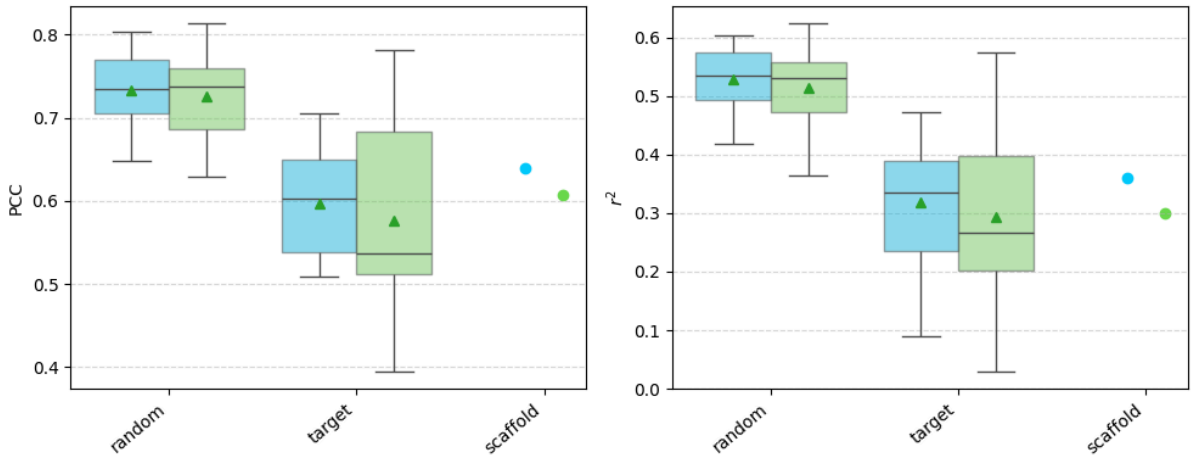

Figure 7: Affinity prediction performance of models trained on crystallographic-only data and models augmented with MD data. RF model trained on static descriptors (derived from crystallographic data: green), and augmented with dynamic and time series descriptors (derived from MD simulations: red). Triangle: mean; horizontal line inside box: median. Scaffold split for a given ligand dataset is deterministic in nature, therefore only a single measurement point is visible.

Features of a representation (i.e. descriptors) may have varying impact on many prediction tasks. Here we compare the correlations of each descriptor derived from crystallographic data, and the ts_descriptors derived from MD simulations, to assess how they influence the affinity prediction task. Figure 8 shows that the application of molecular dynamics results in increased correlations within almost all decryptors (with the exception of the Anionic term). In the group of pocket geometric descriptors, the correlation values are comparable. In this case of pocket descriptors, ts_descriptors resulted in multiple correlation increases. For example, the pocket_C descriptor has a Pearson correlation of 0.00, compared to 0.57 obtained by its ts_descriptor: benford_correlation. The individual performance of the ts_descriptors however does not seem to add up to the final model performance (Figure 7). This effect is probably due to non-linear cross-correlations present between ts_descriptors. Such nonlinearity is difficult to filter beforehand. Figure 8 clearly shows that with the right procedure for feature selection, the use of MD can introduce relevant information not present in the crystallographic structures.

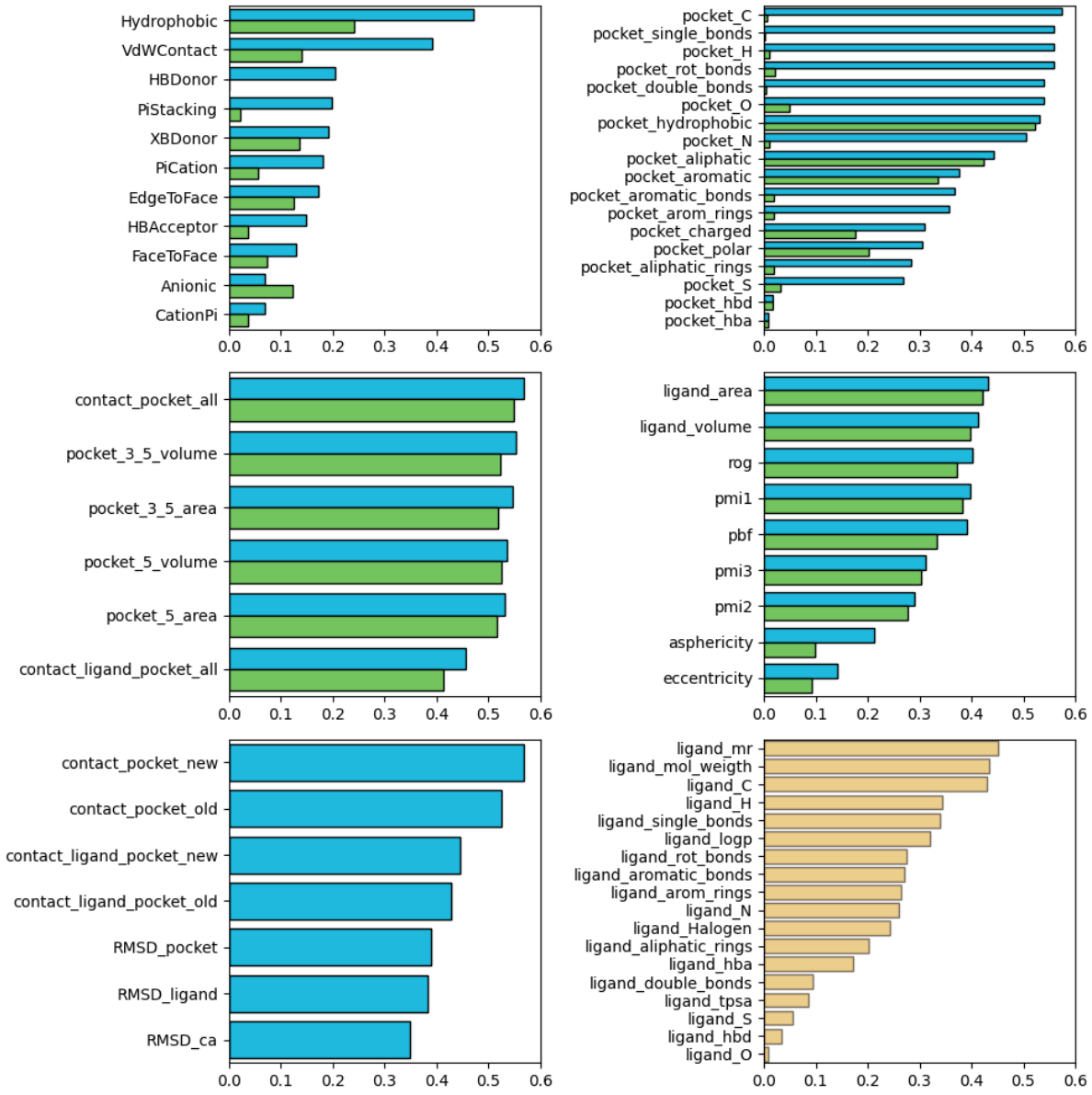

Figure 8: Pearson correlation coefficient of six types of descriptors expressed by the absolute value of the descriptor in relation to affinity. The green color represents the results obtained by descriptors calculated from crystallographic structures, blue with their corresponding ts_descriptors as calculated from the MD simulations (20ns), used in the model.

Ablation studies

Ablation studies involve removing or disabling elements of a model, such as features or parameters, to understand their individual contributions to overall performance or behavior. Table 3 provides a summary of such studies done with both models on two types of splits.
Overall, the MD-derived model performs better in nearly all comparisons. For the target split, the differences between the crystallographic and the MD-derived model performance are higher than for the random split, confirming previously observed better generalizability of the

latter model. It is worth noting that MD-derived model has a lower variance in almost all cases.

The crystallographic model performs very well when trained only on pocket descriptors. However, in the case of a target split, the crystallographic model trained only on the pocket descriptors performs significantly lower than the MD-derived model from which it was superior on the random split. One explanation of this result is the interdependence of pocket and ligand descriptors. Since pocket descriptors are calculated with respect to the ligand position, ligand information is also implicitly contained.

An interesting result was achieved by the models trained only with ECFP4 and ligand properties. These models have an elevated performance in affinity prediction, suggesting they learn certain biases and random relationships in the data rather than predict affinity as a function of both target and ligand complex. Similar conclusions in the context of protein-ligand affinity have been noted in other work as well [45–47]

	Random Split				Target split			
	PCC		r2		PCC		r2	
Descriptors	MD	Crystal	MD	Crystal	MD	Crystal	MD	Crystal
Base (all desc.)	**0.733 (0.045)**	0.727 (0.050)	**0.529 (0.057)**	0.513 (0.064)	**0.598 (0.066)**	0.576 (0.102)	**0.318 (0.104)**	0.294 (0.136)
ligand prop. + ECFP4	0.691 (0.036)	0.691 (0.036)	0.463 (0.044)	0.463 (0.044)	0.522 (0.088)	0.522 (0.088)	0.212 (0.134)	0.212 (0.134)
pocket desc.	0.679 (0.043)	**0.714 (0.044)**	0.453 (0.054)	**0.501 (0.058)**	**0.544 (0.074)**	0.524 (0.073)	**0.257 (0.103)**	0.229 (0.103)
motion desc.	0.582 (0.051)	-	0.331 (0.058)	-	0.532 (0.073)	-	0.239 (0.103)	-
Interaction desc.	**0.509 (0.050)**	0.463 (0.059)	**0.250 (0.050)**	0.196 (0.052)	**0.450 (0.084)**	0.257 (0.141)	**0.147 (0.107)**	-0.01 (0.103)
ligand geometry	**0.507 (0.060)**	0.493 (0.061)	**0.249 (0.061)**	0.230 (0.063)	**0.437 (0.114)**	0.417 (0.106)	**0.119 (0.171)**	0.100 (0.152)
pocket geometry	**0.585 (0.044)**	0.583 (0.054)	**0.334 (0.050)**	0.329 (0.066)	**0.531 (0.072)**	0.526 (0.083)	**0.234 (0.088)**	0.225 (0.124)

Table 4. Ablation studies of the crystallographic and MD-derived models. The 'Descriptors' column describes the sole group of descriptors on which the model has been trained. Standard deviation values are given in brackets.

Conclusion

The application of molecular dynamics in the context of protein ligand affinity prediction may offer advantages. One of the important conclusions derived from this work is that the length of the MD simulation did not substantially improve affinity prediction (Figure 7). The results

show that within the tested length (10-200ns), there is a slight performance gain from 10ns to 20ns, both for random and target splits, however conducting more lengthy simulations did not prove useful for the predictive models. In addition, when correlating time series descriptors to affinity, significantly higher Pearson correlation values for multiple descriptors were observed for the 20ns model compared to 200ns. One possible explanation is longer simulations also introduce more variational noise, difficult to filter by the simple ML models. This conclusion might change however with longer timescales, other types of targets or setups but when testing narrow time windows (under 0,2ps), a short simulation is enough to extract relevant dynamic features. This conclusion brings hope to the inclusion of MD simulation into protocols concerning diverse chemical library screening and hit prioritization and is also consistent with some previous works done on specific targets [10,16,17].

It is important to note the observed better generalization potential of the MD augmented models (Figure 7). In this context generalization refers to the ability of a model to perform well on unknown targets, not present in the training dataset. In this work this is ensured by splitting the data using Uniprot ID rather than PDBID. Indeed for such challenging splits the models trained on MD-derived descriptors show performance advantage and importantly a substantially lower variance in the form of a smaller SD. Moreover the better generalization potential is obtained despite a much lower number of MD training examples compared to crystallographic examples. The difference is not observed with randomly split data, where both models achieve comparable performance (Figure 7).

In the case of novel feature representation, the two models consider different descriptors to be most relevant; ligand descriptors for the crystallographic model and the descriptors directly derived from motion for the MD-derived model. Interestingly, we noted a number of time series derived descriptores with significantly better correlations compared to their static counterparts (Figure 8). Their summarized influence however did transfer only slightly to improved affinity prediction performance. Given the cross-correlation filtering was conducted, this would suggest a lot of non-linear dependence decreasing the overall performance. The ablation analysis and Shap studies further confirms these findings, showing the two models employ different types of descriptors. Despite the difference, results of both models are quite similar for random splits. However for the target splits, an advantage of MD-derived models can be observed. This may indicate generalizability advantages of the time series descriptors, further highlighting their potential for further application.

In conclusion, we have developed the largest, publicly available dataset of molecular dynamics simulations of protein ligand complexes simulations. We treat the MD-derived data as time series in order to extract meaningful statistics and other characteristics of the data that would improve the task of affinity prediction. We found that using simple ML models in combination with a relatively small number of descriptors yields results comparable to highly complex models based on neural networks, with a large number of parameters. We highlight that short (~10ns) molecular dynamics simulations provide relevant information for affinity prediction and that elongation of the simulations does not improve predictive power. Finally, we conclude that models based on MD-derived data treated as time series do not achieve significantly better results compared to models based on crystallographic data, although they appear to be better at generalization. It should be emphasized that the analyzed set, despite being the largest in the currently available literature, is still relatively small and the conclusions should be tested against a larger body of structures.

Funding

This work was sponsored by grant 2020/39/B/ST4/02747 obtained from the Polish National Science Center. Computational resources were provided partially by POL-OPENSCREEN HE ERIC project.

Data Availability

Descriptor generation scripts are available in the repository:
https://github.com/JPoziemski/md_for_affinity_prediction.
Trajectories of molecular dynamics are deposited at: https://zenodo.org/records/11172815.
Other data can be made available upon request.

References

1. Kairys V, Baranauskiene L, Kazlauskiene M, Matulis D, Kazlauskas E. Binding affinity in drug design: experimental and computational techniques. Expert Opin Drug Discov. 2019;14: 755–768.

2. Mobley DL, Gilson MK. Predicting Binding Free Energies: Frontiers and Benchmarks. Annu Rev Biophys. 2017;46: 531–558.

3. Parks CD, Gaieb Z, Chiu M, Yang H, Shao C, Walters WP, et al. D3R grand challenge 4: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. J Comput Aided Mol Des. 2020;34: 99–119.

4. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics. 2010;26: 1169–1175.

5. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. Bioinformatics. 2018. doi:10.1093/bioinformatics/bty374

6. Wang Z, Zheng L, Liu Y, Qu Y, Li Y-Q, Zhao M, et al. OnionNet-2: A Convolutional Neural Network Model for Predicting Protein-Ligand Binding Affinity Based on Residue-Atom Contacting Shells. Front Chem. 2021;9: 753002.

7. Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative Assessment of Scoring Functions: The CASF-2016 Update. J Chem Inf Model. 2019;59: 895–913.

8. Shen C, Hu Y, Wang Z, Zhang X, Zhong H, Wang G, et al. Can machine learning consistently improve the scoring power of classical scoring functions? Insights into the role of machine learning in scoring functions. Brief Bioinform. 2021;22: 497–514.

9. Ganesan A, Coote ML, Barakat K. Molecular dynamics-driven drug discovery: leaping forward with confidence. Drug Discov Today. 2017;22: 249–269.

10. Gu S, Shen C, Yu J, Zhao H, Liu H, Liu L, et al. Can molecular dynamics simulations

improve predictions of protein-ligand binding affinity with machine learning? Brief Bioinform. 2023;24. doi:10.1093/bib/bbad008

11. Gioia D, Bertazzo M, Recanatini M, Masetti M, Cavalli A. Dynamic Docking: A Paradigm Shift in Computational Drug Discovery. Molecules. 2017;22. doi:10.3390/molecules22112029

12. Śledź P, Caflisch A. Protein structure-based drug design: from docking to molecular dynamics. Curr Opin Struct Biol. 2018;48: 93–102.

13. Guterres H, Im W. Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations. J Chem Inf Model. 2020;60: 2189–2198.

14. Riniker S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. J Chem Inf Model. 2017;57: 726–741.

15. Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des. 2014;28: 711–720.

16. Ash J, Fourches D. Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories. J Chem Inf Model. 2017;57: 1286–1299.

17. Jamal S, Grover A, Grover S. Machine Learning From Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors Against Alzheimer's Disease. Front Pharmacol. 2019;10: 780.

18. Kyaw Zin PP, Borrel A, Fourches D. Benchmarking 2D/3D/MD-QSAR Models for Imatinib Derivatives: How Far Can We Predict? J Chem Inf Model. 2020;60: 3342–3360.

19. Liu Z, Su M, Han L, Liu J, Yang Q, Li Y, et al. Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. Acc Chem Res. 2017;50: 302–309.

20. Simonovsky M, Meyers J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. J Chem Inf Model. 2020;60: 2356–2366.

21. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol. 2017;13: e1005659.

22. Kagami L, Wilter A, Diaz A, Vranken W. The ACPYPE web server for small-molecule MD topology generation. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btad350

23. Bayly CI, Cieplak P, Cornell W, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J Phys Chem. 1993;97: 10269–10280.

24. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1-2: 19–25.

25. RDKit: Open-source cheminformatics. Available: http://www.rdkit.org

26. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020;17: 261–272.

27. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. J Comput Chem. 2011;32: 2319–2327.

28. Bouysset C, Fiorucci S. ProLIF: a library to encode molecular interactions as fingerprints. J Cheminform. 2021;13: 72.

29. Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing. 2018;307: 72–77.

30. Ramsundar B, Eastman P, Walters P, Pande V, Leswing K, Wu Z. Deep Learning for the Life Sciences. O'Reilly Media; 2019.

31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

32. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1603.02754

33. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Min Knowl Discov. 2014;28: 92–122.

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

35. Durrant JD, McCammon JA. NNScore 2.0: a neural-network receptor-ligand scoring function. J Chem Inf Model. 2011;51: 2897–2903.

36. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31: 455–461.

37. Cang Z, Mu L, Wei G-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. PLoS Comput Biol. 2018;14: e1005929.

38. Zhang S, Jin Y, Liu T, Wang Q, Zhang Z, Zhao S, et al. SS-GNN: A Simple-Structured Graph Neural Network for Affinity Prediction. ACS Omega. 2023;8: 22496–22507.

39. Liu X, Feng H, Wu J, Xia K. Dowker complex based machine learning (DCML) models for protein-ligand binding affinity prediction. PLoS Comput Biol. 2022;18: e1009943.

40. Wee J, Xia K. Ollivier Persistent Ricci Curvature-Based Machine Learning for the Protein-Ligand Binding Affinity Prediction. J Chem Inf Model. 2021;61: 1617–1626.

41. Zhang X, Gao H, Wang H, Chen Z, Zhang Z, Chen X, et al. PLANET: A Multi-objective Graph Neural Network Model for Protein-Ligand Binding Affinity Prediction. J Chem Inf Model. 2023. doi:10.1021/acs.jcim.3c00253

42. 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. 2018;58: 287–296.

43. Zheng L, Fan J, Mu Y. OnionNet: a multiple-layer inter-molecular contact based convolutional neural network for protein-ligand binding affinity prediction. arXiv [physics.bio-ph]. 2019. Available: http://arxiv.org/abs/1906.02418

44. Shen C, Zhang X, Hsieh C-Y, Deng Y, Wang D, Xu L, et al. A generalized protein-ligand scoring framework with balanced scoring, docking, ranking and screening powers. Chem Sci. 2023;14: 8129–8146.

45. 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. 2019;59: 947–961.

46. Volkov M, Turk J-A, Drizard N, Martin N, Hoffmann B, Gaston-Mathé Y, et al. On the Frustration to Predict Binding Affinities from Protein-Ligand Structures with Deep Neural Networks. J Med Chem. 2022;65: 7946–7958.

47. Libouban P-Y, Aci-Sèche S, Gómez-Tamayo JC, Tresadern G, Bonnet P. The Impact of Data on Structure-Based Binding Affinity Predictions Using Deep Neural Networks. Int J Mol Sci. 2023;24. doi:10.3390/ijms242216120