Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5dd00452g

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

Jakub Poziemski,[a] Artur Yurkevych[b] and Pawel Siedlecki *[a]

Received 7th October 2025
Accepted 18th March 2026

DOI: 10.1039/d5dd00452g

rsc.li/digitaldiscovery

The advancements in computational methods in drug discovery, particularly through the use of machine learning (ML) and deep learning (DL), have significantly enhanced the precision of binding affinity predictions. However, accurate prediction of binding affinity remains a challenge due to the complex, non-linear character of molecular interactions. Generalizability continues to limit the current 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 their predictive performance and limitations. In particular, MD simulations offer a dynamic perspective by depicting the temporal interactions within protein–ligand complexes, potentially providing 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 accuracy. The findings suggest specific and generalizable features derived from MD data and propose approaches to augment the current *in silico* affinity prediction methods.

Introduction

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.[1] 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] Progress in machine learning (ML) and deep learning (DL) has shown promise in overcoming some of these hurdles,[4–6] revealing intricate, non-linear properties and relationships in protein–ligand complexes in large datasets. Current state-of-the-art methods achieve a Pearson correlation coefficient (PCC) of around 0.7–0.85 on the CASF2016 benchmark,[7,8] which is a significant improvement compared to the classical scoring functions. Despite this achievement, challenges remain, particularly with the generalizability of these models. While definitely useful, traditional static computational approaches like molecular docking only provide a snapshot view of a molecular complex, without its temporal dynamics.[9,10]

Molecular dynamics (MD) simulations introduce a vital temporal dimension to protein–ligand complex studies. Such simulations allow for more detailed observations of how drug molecules interact with biological targets over time.[11–13] Over the last few years, integration of MD simulations with ML and DL has been applied with varying success in different drug discovery tasks and specific campaigns. In the case of affinity prediction, Ash and Fourches in 2017 (ref. 14) analyzed 87 ERK2- docked ligand complexes by computing chemical descriptors derived from 20 ns 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 library screening tasks, candidate design and lead prioritization. A similar conclusion was presented by the authors of ref. 15, who performed molecular docking of 43 compounds associated with Caspase-8, with consecutive 10 ns MD simulations of the 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 (R_g), 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 [16] using the BCR-ABL tyrosine-kinase and 15 ns MD simulations of Imatinib and a large series of its derivatives. In conclusion the authors stated that incorporating MD based matrices could not improve the binding affinity prediction ability of either deep NN or random forest (RF) QSAR models. However, their models did show

[a]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland. E-mail: pawel@ibb.waw.pl
[b]Institute of Chemistry, University of Silesia in Katowice, Katowice, Poland

reduced prediction error, indicating that the negative effect of simulation noise becomes stronger as the number of snapshots increased. Another approach to compare different ML models trained on descriptors obtained from MD trajectories was presented in ref. 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 on the target. The paper concludes that the use of MD does not generally improve screening results and may only be justified in certain cases. Given that the models were trained with descriptors generated from every frame, this may have been a challenge for simple ML models due to the large number of frames and small number of examples. In addition, the low MM/PBSA and Glide scores suggest that the analyzed collections were rather difficult. In conclusion, current research indicates the complexity of leveraging molecular dynamics (MD) data, suggesting that 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.

In this study, we have generated the largest set of MD simulations to date, encompassing a broad array of protein–ligand complexes. We developed 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 time series descriptors from over 800 unique protein–ligand complexes, we seek answers to the following questions: (1) what are the complex specific and simulation specific features influencing the affinity prediction outcomes of the models? (2) Whether time series representations of MD are beneficial and how do they generalize? (3) Can the MD-derived descriptors augment and/or replace current crystallographic derived descriptors? Based on our findings, we present general guidelines and assessments on how such an approach influences the *in silico* affinity prediction on a large scale.

Methods

Dataset compilation

The Molecular Dynamics Dataset (MDD) comprises 231 unique targets (Fig. 1) from 862 protein–ligand complexes, sourced from the PDBBind collection v2020.[19] Only complexes with well-defined active sites and ligands with unambiguous affinity values ($-\log K_i$, K_d and IC_{50}) were considered. More details on the filtering procedure, functional composition and similarity assessments of MDD targets and ligands are available in the SI.

MD simulation procedure

The MD preparation protocol, identical for all MDD complexes, is detailed in the SI. Briefly, MD simulations were executed using GROMACS,[17] utilizing a cubic box under periodic boundary conditions and a TIP3P water model. An initial minimization cycle, followed by temperature equilibration in the *NVT* ensemble and pressure equilibration in the *NPT* ensemble was carried out. Production simulations were conducted over a 200 ns timeframe, with a timestep of 100 ps.

Representation

Given the relatively small size of the MDD compared to *e.g.* PDBBind, we limit the number of descriptors to minimize overfitting, data sparsity and avoid other unfavorable phenomena caused by the curse of dimensionality.[18]

The crystallographic (static) representation is composed of 63 descriptors: 24 pocket descriptors calculated with RDKIT and MDAnalysis[19,20], 11 interaction descriptors calculated with ProLIF[21] and 28 ligand descriptors calculated with RDKIT and SciPy[19,22] (for details see the "List of descriptors" section in the SI).

In the case of MD simulation data, for each complex we calculate 51 descriptors: 24 pocket descriptors, 11 interaction descriptors, 9 ligand descriptors (19 ligand property descriptors are omitted as they do not change during simulation), and additional 7 motion descriptors not present in static representations. These 51 descriptors are calculated for each frame of the MD simulation. For each descriptor sequence, we extract its unique time series value with the use of a multistep procedure:

(1) For each descriptor, we calculate all 788 time series descriptors (ts_descriptors) supported by tsfresh (see Fig. 2).

(2) For each ts_descriptor, its *p*-value is determined in terms of statistical significance against experimentally determined affinity (sourced from PDBBind), using a univariate test with FDR set to 0.001.

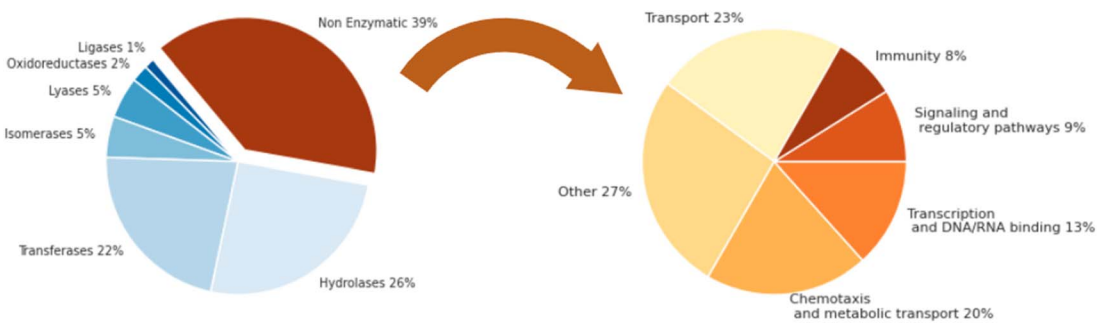

Fig. 1 Functional characterization of MDD targets. The "non-enzymatic" part of the MDD ($\frac{1}{3}$ of the targets) is described on the right chart by 5 distinct biological processes (GO annotation). Nearly 40% of all MDD targets are non-enzymatic proteins with other functions.

(3) Next, a trimming procedure is used. For each of the 75 feature types,[23] the ts_descriptor with the lowest *p*-value is selected. Some ts_descriptors are parametric in nature; for such cases, we use different thresholds to generate several versions of that ts_descriptor and select the one with the lowest *p*-value. At this stage, there could be a maximum of 75 ts_descriptors per single descriptor.

(4) To avoid caveats in training, trimming of correlations 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 retained.

(5) In the final filtering step, for each descriptor, the ts_descriptor with the highest mutual information score between

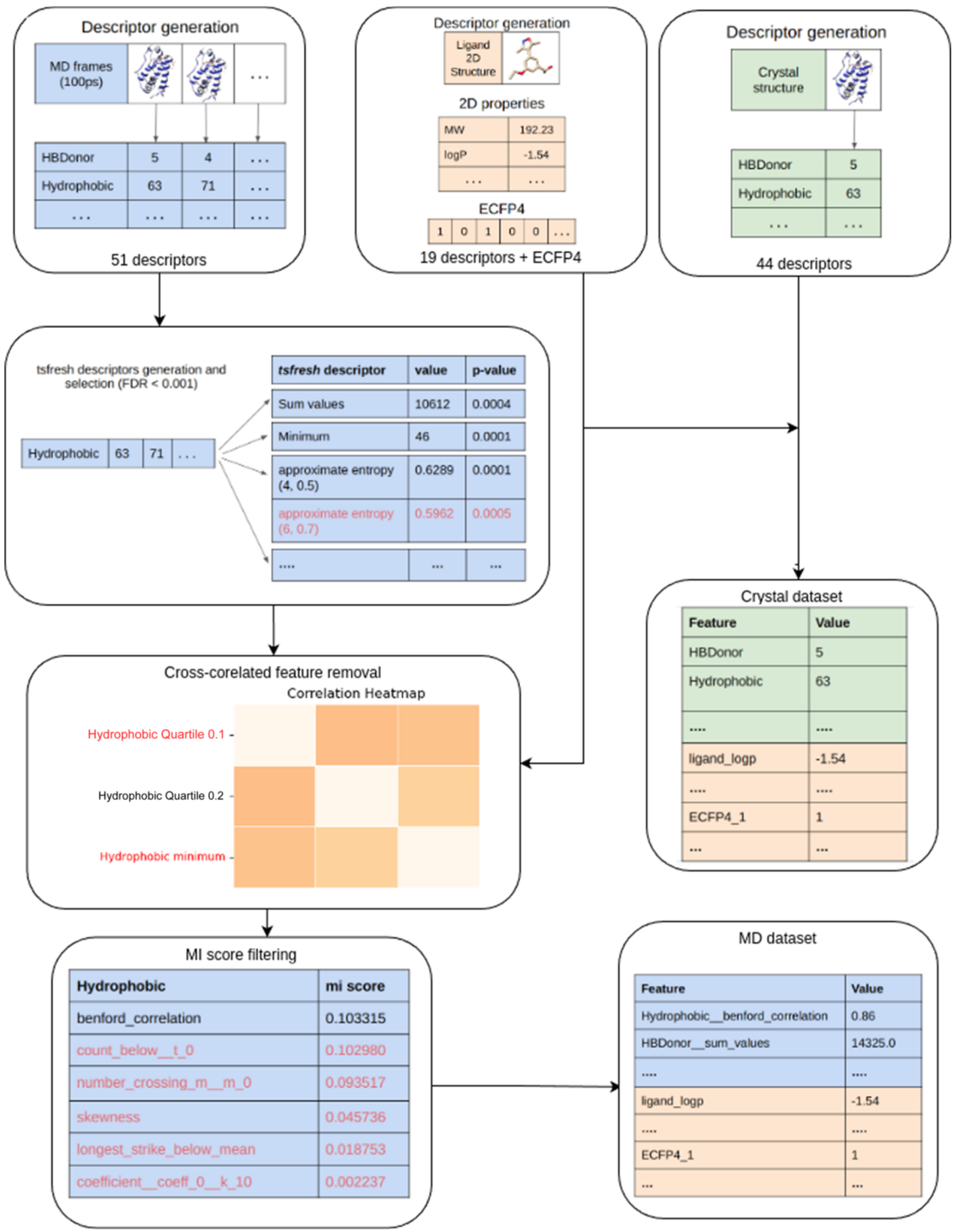

Fig. 2 Flowchart of descriptors and the ts_descriptor selection procedure. Color code: light green – crystallographic (static) descriptors, light blue – multiframe MD descriptors and ts_descriptors, and light orange – ligand descriptors. See section "Representation" for more details.

itself and the experimentally determined affinity value was selected. As a result, each descriptor is described with at most one ts_descriptor (Fig. 2, bottom left).

In total, each crystallographic complex is described by 63 descriptors. For the MD data representation the same complex is described by a maximum of 51 ts_descriptors and 19 ligand property (static) descriptors, for which the time series cannot be generated. Both representations are augmented with the ECFP4 fingerprint (1024 bits) to directly incorporate molecular connectivity information of ligands.

Data splits, training and testing

Two ways of splitting the target data were used: random split and target split at a ratio of 4 : 1 (80% training collection and 20% test collection). In the random split, complexes were randomly allocated to the training and test sets. In the target split, Uniprot IDs split the complexes, therefore, there are no identical training and testing examples. The target split can therefore be used to approximate the generalization potential.

Ligands were split using DeepChem.[24] Scaffold Splitter divides molecules into groups based on their Bemis-Murcko scaffolds; the smallest scaffold groups form the test set. Although not without its caveats, this type of ligand split is more challenging than random splits, as it tests more thoroughly the generalizability to new or less abundant areas of chemical space.

Three different ML models were trained using the MDD: Random Forest, SVM[25] and XGBoost.[26] Model parameters were selected using 5-fold cross-validation (CV). The models were fitted to the training sets and evaluated on the test sets. Throughout this work, boxplots represent results obtained on the test set, with mean (triangle) and median (horizontal line inside the boxplot) values. More details on the Random Forest models and on the descriptor model (XGB) are presented in the SI.

Results

Baseline performance

To test the hypothesis that time series descriptors improve affinity prediction, we first assess the difficulty of predicting the affinity of MDD complexes using published models with publicly available code and training procedures.[27] Selected models were trained on the PDBBind dataset (v2020) with 862 MDD complexes excluded (Table 1). We compare these results against those obtained for the CASF2016 dataset presented in the literature (Table 2).

Both Pearson Correlation Coefficient (PCC) and root mean square error (RMSE) values (Table 1) render the MDD a more difficult dataset compared to CASF2016 (Table 2). All tested models show a rather consistent drop in performance, independent of their complexity. The observed decrease in performance may be multifaceted: CASF2016 is a relatively small dataset compared to MDD (285 *vs.* 862 complexes), therefore, it may be easier to optimize or overtrain the models. Also, excluding MDD complexes from the training data may have influenced the availability of information necessary for higher affinity prediction performance.

Our descriptor model (see the Representation section in the Materials and methods section) shows the same consistent performance drop as seen with other models. Interestingly, the affinity prediction performance of our models is on par with some of the best, highly sophisticated methods. This result highlights that a carefully selected set of descriptors and a relatively simple ML model can show a level of performance comparable to specialized neural networks.

Simulation length

To determine the optimal length of MD simulations for information extraction, we tested 5 timescales from 10 ns up to 200 ns. The results (Fig. 3) shows the performance of the Random Forest models across the tested MD simulation lengths. The results indicate varying correlations for both random and target splits. For random splits, 50 ns runs have slightly higher correlation values (PCC and r2) but show a lower RMSE at 20 ns. For the target split, 20 ns trajectories show the best performance estimates for all three measures (PCC, r^2 and RMSE values). Taken together, the 20 ns simulation length should provide a good performance balance for both random and target splits.

We assessed the difference in affinity correlations between individual time series descriptors (ts_descriptors) derived from 20 ns compared to 200 ns MD simulations. We filtered all cross-

Table 1 Performance of selected affinity prediction methods on the MDD. All models were trained on PDBBind v2020 complexes with MDD complexes excluded. Both PCC (Pearson correlation coefficient) and RMSE (root mean square error) values were calculated for MDD complexes. The training size column shows the number of unique protein–ligand complexes used for training the ML/DL models; according to their original implementation either with the general set or refined set; Vina uses a classical hand-designed scoring function with optimized parameters

Model name	Description	PCC	RMSE	Training size	Code and feature reference
OnionNet2	CNN trained on contact descriptors	0.75	1.26	13 546	6
PLEC-NN	Extended connectivity FP & neural network	0.74	1.54	11 203	28
Descriptor model (XGB)	XGBoost trained on 63 descriptors + ECFP4	0.72	1.29	12 349	This work
NN-score	Feed-forward neural network	0.67	1.40	4647	29
RF-score v3	Random forest with spatial distance count	0.65	1.45	4647	4
Vina	Hybrid empirical scoring function	0.49	—	—	30

Table 2 Published affinity prediction performance obtained with models of increasing complexity, tested with CASF_2016 and CoreSet_2013 benchmarks. PCC (Pearson correlation coefficient), RMSE (root mean square error), training size (number of unique protein–ligand complexes used for training ML/DL models). The use of simple models on crystallographic data can yield comparable results to the use of complex neural network based models

		PCC	RMSE		
Model name	Description	CASF|CoreSet	CASF|CoreSet	Training size	Original reference
OnionNet2	CNN trained on contact descriptors	0.86|0.82	1.16|1.36	—	6
TopBP	Topological descriptors with GBT	0.86|0.81	1.19|1.95	3767	31
SS-GNN	Graph neural network	0.85|0.82	1.18|1.35	15 394	32
Descriptor model (XGB)	XGBoost	0.85|0.81	1.20|1.37	12 866	This work
EBA-AY	Ensemble attention based	0.86|0.79	1.20|1.44	10 324	33
OPRC-GBT	Ollivier persistent Ricci curvature	0.84|0.79	1.25|2.01	3772	34
DCML	Dowker complex based machine learning	0.84|0.78	1.25|1.43	3772	35
CAPLA	Sequence based cross-attention with 1D CNN	0.84| 0.77	1.2|1.36	11 906	36
PLANET	Graph neural network	0.82|N/A	1.24|N/A	15 616	37
K_{DEEP}	Convolutional neural network	0.82|N/A	1.27|N/A	3767	38
PLEC-NN	Extended connectivity FP & neural network	0.82|0.77	1.25|1.43	12 906	28
OnionNet	Convolutional neural network	0.82|0.78	1.27|1.50	11 906	39
RF-score v3	Random forest	0.80|0.74	1.39|1.51	3767	4
Pafnucy	Convolutional neural network	0.78|0.70	1.42|1.62	11 906	5
Vina	Hybrid empirical scoring function	0.60|0.56	1.75|1.86	—	40

correlated ts_descriptors from the two simulation lengths, and compared the shared 36 ts_descriptors. The results show a significant gain in around 40% of the tested ts_descriptors (14 out of 36, with $\Delta > 0.1$) in favor of the 20 ns simulation length. Comparable correlations are registered for 22 ts_descriptors ($\Delta < 0.1$). Interestingly, there was no descriptor that had more than 0.1 correlation difference in favor of the 200 ns simulation.

In the assessed timescale, the obtained results indicate that short MD simulations may capture useful steric conformational changes and that longer MD simulations may contain more random movements or noise, which may lower the individual correlation of some of the ts_descriptors. Similar results have been presented in studies concerning single targets.[10,14,15] Taken together, our results show that time series descriptors from longer MD do not provide an advantage over short MD runs, possibly due to higher noise content and increased noise fitting during model training (Fig. 4).

We also tested whether the frequency of saving the trajectory frames had an influence on model performance. We observed no significant improvement in results when training on an extensive number of frames (see SI Fig. S3). From a practical point, in similar setups we recommend using a less frequent recording (larger time interval), which can significantly reduce storage requirements without decreasing prediction quality.

MD representation

To estimate the impact of time series descriptors on affinity prediction, we compared the overall performance of two types of models: trained with crystallographic descriptors and with MD derived ts_descriptors, with respect to different target and ligand data splits. The results of these experiments are summarized in Fig. 5. All analyses refer to the RF models trained on 20 ns molecular dynamics runs. For randomly split data, both models achieve comparable results with respect to PCC (0.73 *vs.* 0.73), r^2(0.53 *vs.* 0.51) and SD (0.06 *vs.* 0.06). However, in the case of a more challenging target split, an advantage of the model trained on time series descriptors can be seen (PCC mean: 0.6 *vs.* 0.58 and median 0.6 *vs.* 0.53), along with a smaller SD (0.07 *vs.* 0.10). This slight advantage can also be seen with respect to the Bemis-Murcko scaffold split. Both target and scaffold splits are more challenging tests, as they try to minimize data leakage events. For the scaffold split, the MD-based model also achieved better results (PCC 0.63 *vs.* 0.60 and r2 0.35 *vs.* 0.29). Taken together, the performance gain and the lower SD of models trained on time series descriptors would suggest better generalization potential, also with respect to uncharted chemical space.

Next, we compared each crystallographic descriptor with its time series counterpart to assess their correlations with affinity. Fig. 6 shows the results obtained for the 6 different groups of descriptors used. In the group of interaction descriptors, all time series descriptors have increased correlations compared to their crystallographic counterparts (with the exception of the anionic term). For both types of descriptors, the hydrophobic and vdW interactions show the highest affinity correlation, and importantly the time series correlation at least doubles compared to crystallographic descriptors. In the case of a single descriptor, a correlation around 0.5 PCC would be considered fairly high. The correlations for the rest of the time series interaction descriptors are higher than those of crystallographic descriptors; nevertheless, overall their correlation values are low for both types of data.

With respect to pocket descriptors, both property and geometric, the results show a substantial number of them with

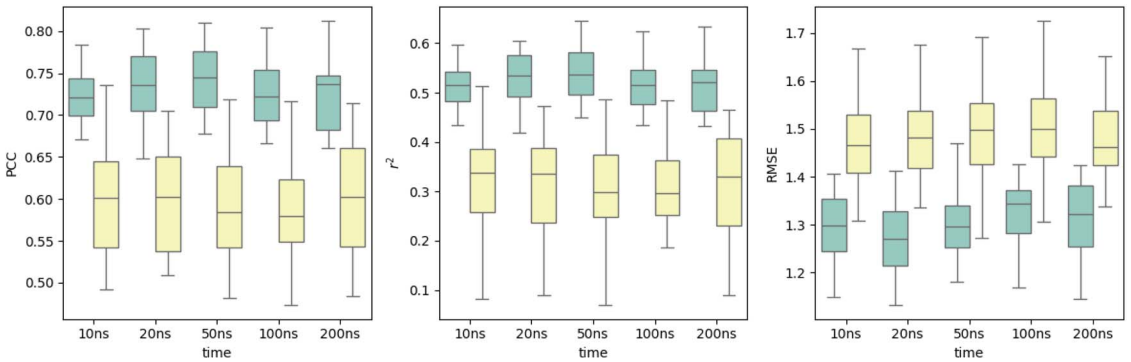

Fig. 3 Dependence of MD simulation length on model performance. RF models trained with trajectories of different lengths (from 10 ns to 200 ns) tested on two types of data splits: random (green) and target (yellow). 20 ns trajectories show good overall performance for both random and target splits.

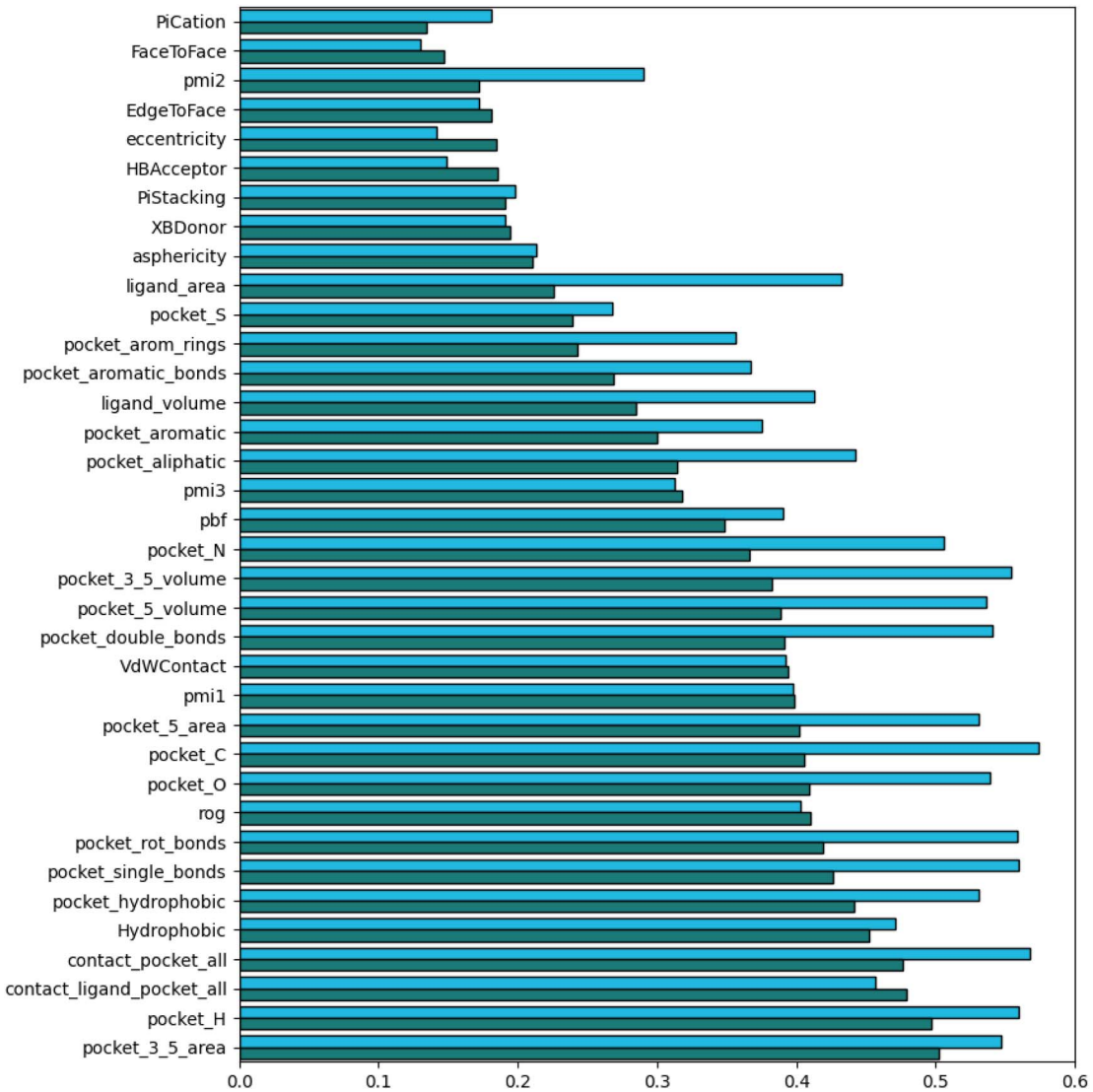

Fig. 4 Correlations between affinity prediction and time series descriptors describing the MDD protein–ligand complexes. ts_descriptors derived from 20 ns (light blue) and 200 ns (dark blue) molecular dynamics simulations.

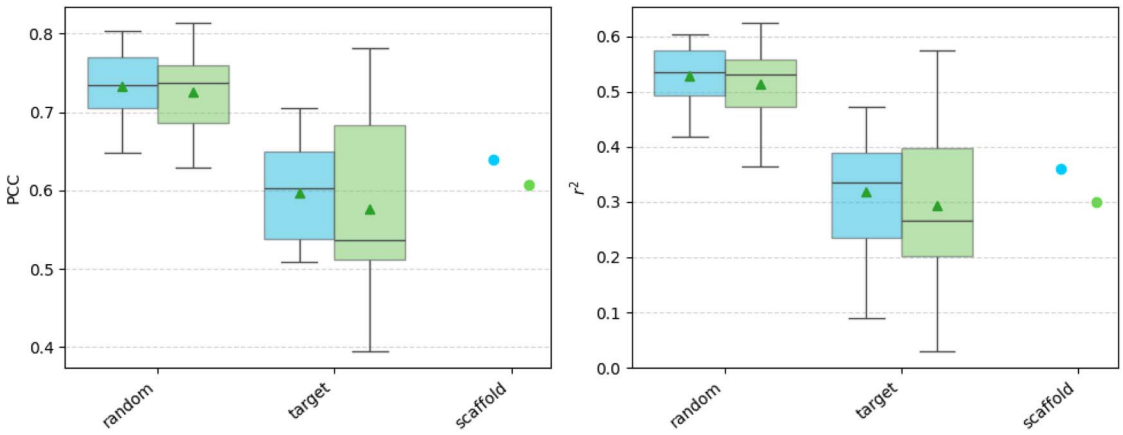

Fig. 5 Affinity prediction performance of models trained on crystallographic-only data and models augmented with MD data. RF models trained on static descriptors (crystallographic data: green), and time series descriptors (MD simulations: blue). Triangle: mean; horizontal line inside the box: median. Scaffold split for a given ligand dataset is deterministic in nature, therefore only a single measurement point is visible.

PCC close or above 0.5. Overall, pocket property descriptors show the highest difference between static and dynamic treatment. Out of the 8 best correlating time series descriptors only one static descriptor (pocket_hydrophobic) has a comparable correlation. The results show that simple pocket composition features, such as atom or bond types count, which change with respect to ligand movement, correlate with high PCC values with small molecule affinity.

In the case of geometric descriptors, correlation values are comparable between time series and static treatment. Similar results are obtained for ligand geometric descriptors. Here, molecular eccentricity and molecular asphericity descriptor correlations gain the most from a time series representation; however their PCC values are rather low.

Surprisingly, motion descriptors, which we thought would substantially contribute to the affinity prediction performance, show only minor gains compared to static, crystallographic descriptors. Only two of them (pocket_contact_new and pocket_contacts_old) have PCC around 0.5 or above. The results suggest that pocket internal contacts (both preserved and new) are more correlated with affinity than *e.g.* pocket–ligand interactions and changes in RMSD of both molecular entities.

Taken together, although individual performance of the ts_descriptors is favorable, it does not seem to add up to the final model performance (Fig. 5). This effect is probably due to non-linear cross-correlations present between the ts_descriptors. This effect is further enhanced by the noise in the data. Common correlation methods such as Pearson's, Spearman's or Kendall are able to find linear and monotonic relationships between variables, assuming little noise. Other methods measuring potential non-linear correlations test only the independence of variables without quantifying the strength of the relationship, and are also very sensitive to noise and outliers. Therefore, such nonlinearity is currently difficult to filter beforehand.[41]

Screening and pose selection

We tested the impact of models trained with molecular dynamics (MD) data and with static crystallographic structures on a screening campaign using one of the targets derived from the DUD-E dataset.[42] The glucocorticoid receptor (GCR, PDB ID: 3BQD) was selected along with a randomly drawn subset of 100 ligands; comprising 10 active molecules and 90 decoys. This composition mimics a virtual screening scenario with a high imbalance between actives and inactives. Each GCR–ligand complex was docked using Vina with default parameters, except that exhaustiveness was set to 32. Next, the best scoring conformation was subjected to a 20 ns MD simulation using the same parametrization procedure as for all other complexes (see the Materials and methods section). Table 3 summarizes the results of GCR screening.

The MD-derived model showed better performance in early enrichment (EF10%) compared to the static model (2.82 *vs.* 1.92). This result was accompanied by a higher ROC AUC score (0.675 *vs.* 0.615). Overall the MD-derived model improved the ability to discriminate between active and inactive compounds, indicating a potential benefit of using the dynamic information in the training process.

To evaluate the ability of the two models to select the correct ligand binding pose, we conducted an experiment using a subset of targets derived from the CASF-2016 dataset. We selected 3 targets: FAX (1lpg, 1z6e, and 2xbv), CDK2 (1pxn and 4eor) and CAH (2weg and 3dd0). For each of the 7 target–ligand complexes, a docking run was performed following the same procedure as described for the screening experiment (*i.e.* Vina, default parameters, exhaustiveness = 32). To minimize bias, ligand structures used for docking were generated from SMILES strings using Open Babel's lowest energy parameter. Each docking run resulted in nine putative conformations for every protein target. The conformation derived directly from the crystallographic structure was included, resulting in 10 poses per target. Finally, each of the obtained poses were subjected to a 20 ns MD simulation using the same parametrization procedure, to obtain MD-derived descriptors (see the Materials and methods section). The results of the ability of the two types of models to correctly identify the native crystallographic structure (*i.e.* assign the highest affinity value) among the docking results are shown in Table 3.

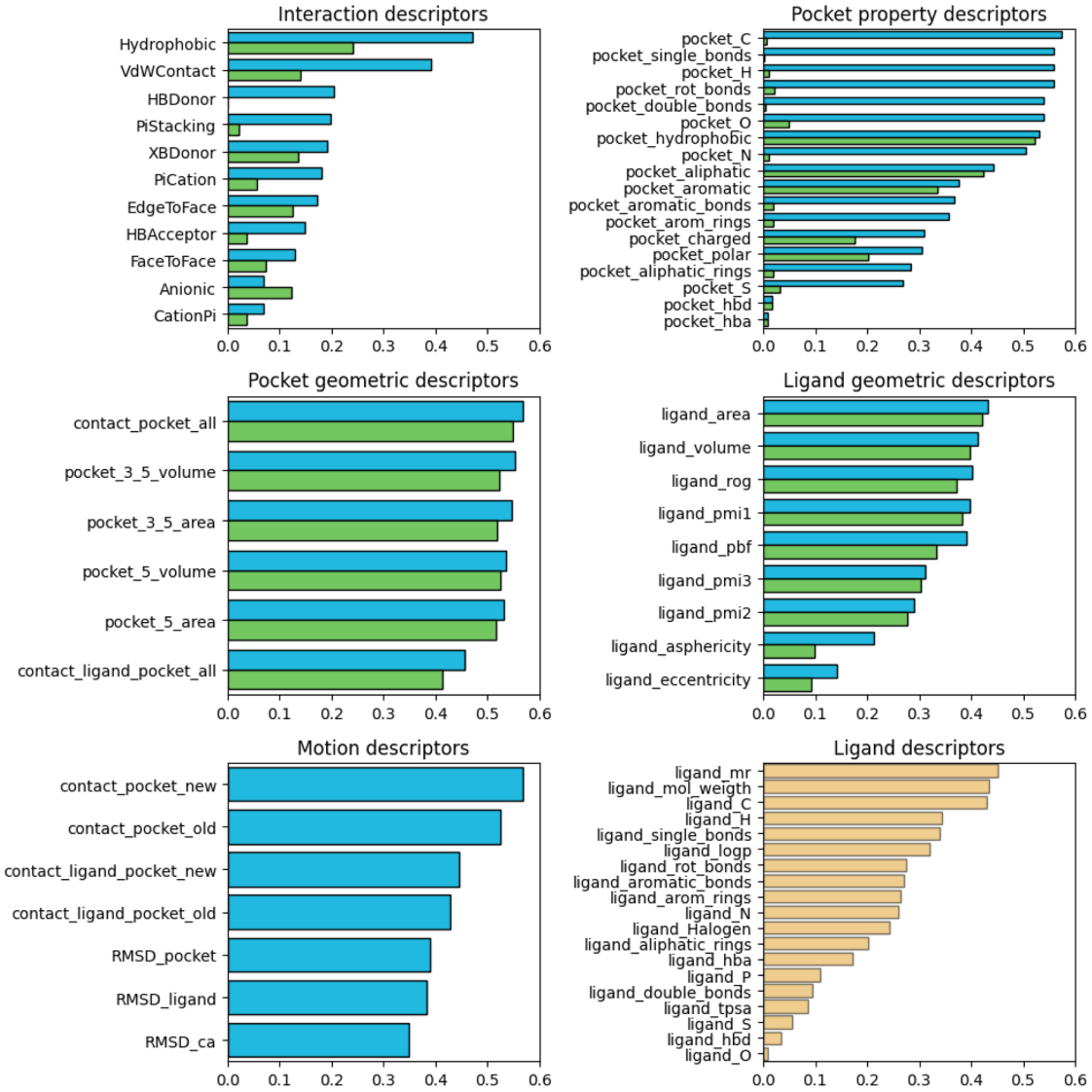

Fig. 6 Pearson correlation coefficient of the six types of descriptors expressed by the absolute value of the descriptor in relation to affinity. Green represents results obtained from descriptors calculated from crystallographic structures, while blue represents their corresponding ts_descriptors as calculated from the MD simulations (20 ns), used in the model.

Table 3 Screening results for the static and MD-derived model (GCR target) and pose selection

	Static model	MD-derived model
Screening task (GCR target)		
EF10%	1.92	2.82
ROC AUC	0.615	0.675
Pose selection (7 target structures)		
Top-1	2/7	4/7
Top-3	3/7	5/7
Mean std	0.22	0.28

The model based on MD trajectories performs better than the model based on crystallographic structures only. However, it is also noticeable that the standard deviation is higher for the MD model (0.28 *vs.* 0.22), suggesting a greater sensitivity of this approach to the choice of initial conformation. This may reflect the complex nature of the molecular dynamics process and indicates that accounting for conformational variation affects the stability of the prediction. Ultimately, these results suggest that models based on MD simulation data may be an effective alternative to approaches based only on static crystallographic structures.

Ablation studies

To understand the contributions of static and time series descriptors to model behavior, we performed ablation studies, including SHAP analysis.[43] Table 4 provides a summary of ablation studies where only a certain group of descriptors or ts_descriptors are used for training. Overall, for the target splits, the time series descriptor based models perform better

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

	Target split				Random split			
	PCC		r^2		PCC		r^2	
Descriptors	MD	Crystal	MD	Crystal	MD	Crystal	MD	Crystal
All descriptors	**0.60 (0.07)**	0.58 (0.10)	**0.32 (0.10)**	0.29 (0.14)	**0.73 (0.04)**	0.73 (0.05)	**0.53 (0.06)**	0.51 (0.06)
Pocket property	**0.54 (0.07)**	0.52 (0.07)	**0.26 (0.10)**	0.23 (0.10)	0.68 (0.04)	**0.71 (0.04)**	0.45 (0.05)	**0.50 (0.060)**
Motion	0.53 (0.07)	—	0.24 (0.10)	—	0.58 (0.05)	—	0.33 (0.06)	—
Pocket geometric	**0.53 (0.07)**	0.53 (0.08)	**0.23 (0.09)**	0.22 (0.12)	**0.58 (0.04)**	0.58 (0.05)	**0.33 (0.05)**	0.33 (0.07)
Ligand property + ECFP4	0.52 (0.09)		0.21 (0.13)		0.69 (0.04)		0.46 (0.04)	
Interaction	**0.45 (0.08)**	0.26 (0.14)	**0.15 (0.11)**	−0.01 (0.10)	**0.51 (0.05)**	0.46 (0.06)	**0.25 (0.05)**	0.20 (0.05)
Ligand geometric	**0.44 (0.11)**	0.42 (0.11)	**0.12 (0.17)**	0.10 (0.15)	**0.51 (0.06)**	0.49 (0.06)	**0.25 (0.06)**	0.23 (0.06)

compared to their static model counterparts; however, we note that the difference is not major. The highest performance was obtained when training exclusively with pocket property ts_descriptors, closely followed by pocket geometry ts_descriptors. These results are in line with Fig. 6, where the time series representation of pocket properties provided substantially more information useful for affinity correlation than a static representation.

Interestingly, for random splits, models trained only on static pocket property descriptors perform better than models trained on their time series counterparts. The situation changes with target splits; the static models perform significantly worse compared to the MD-derived models, to which they were previously superior. 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 implicitly contained, even more so with a dynamic representation of the pocket.

The highest difference between crystallographic and MD derived models is obtained when training with interaction descriptors (0.450 *vs.* 0.257). This result confirms our initial hypothesis that the interaction and motion descriptors when represented as time series provide novel and useful information in the context of affinity prediction. However, these ts_descriptors, when combined with other types, fail to make a substantial difference in performance. This might indicate the need for a more thorough representation of motion and interactions present in ligand–receptor complexes, compared to the setup tested in this work.

An interesting result was achieved by the models trained only with ECFP4 and ligand properties. These models, having no information about the target, show elevated performance in affinity prediction, suggesting that they mostly learn biases and random relationships in the data rather than predict affinity as a function of both the target and ligand complex. Similar conclusions in the context of protein-ligand affinity predictions have been noted in other studies as well.[44–46]

SHAP analysis presents the 20 most important features, together with their utilization in the form of counts, over all 20 models of each type; static- and time series based (Table S7). Each model is using a slightly different set of descriptors (or ts_descriptors) depending on data splits it was trained on. SHAP analysis of a single static- and time series based model is presented in Fig. S4 in the SI as a reference.

One striking difference between the static and time series models is the heavy reliance on ligand descriptors by the former. Out of the 20 most influential descriptors, nearly half of them (9/20, 45%) are ligand based. The same comparison with time series models shows that they rely only on 25% (5/20) of ligand based ts_descriptors. Even more, with static models, 5/9 of the most influential ligand descriptors are simple 2D physicochemical features. With time series models it is just 1/5. These results may explain the poorer performance of static models observed with the scaffold splits.

SHAP analysis shows that the MD-derived models rely mostly on a different set of features compared to models trained on static representations. There are three motion descriptors important for the time series models. The pocket contacts (both old and new) and ligand RMSD are especially interesting, as they are exclusive to the MD-derived representation.

Taken together, the SHAP results show that the time series models use different sets of features, rely on simulation exclusive motion descriptors, and use less simple ligand features, rendering them possibly less prone to small molecule bias present in datasets.

Conclusion

In this work, we introduce the MDD, the largest publicly available collection of 200 ns molecular dynamics simulations of 862 protein–ligand complexes. This resource enables systematic investigation of time-series descriptor extraction strategies, feature-selection protocols, and provides a scalable foundation for extending both the size of training sets and the duration of simulations, thereby supporting the development and benchmarking of next-generation MD aware structure-based models in drug discovery.

The novelty of our approach lies in a feature-centric representation of MD data, where selected protein–ligand interaction properties are tracked over time and summarized using time-series descriptors (ts_descriptors). Unlike snapshot-based augmentation strategies commonly employed in prior studies,

this framework captures ensemble-level interaction tendencies through temporal statistics, enabling efficient learning from MD trajectories using tabular representations.

One of the central findings of this study is that longer MD simulations do not necessarily improve predictive performance. Across more than 800 simulations, models trained on descriptors derived from 20 ns trajectories consistently matched or outperformed those based on 200 ns simulations. This effect is likely attributable to increased variational noise introduced at longer timescales, which is difficult for machine learning models to filter. We further investigated whether the optimal MD simulation length depends on macroscopic protein–ligand properties, such as protein or ligand size or intrinsic flexibility. Despite extensive analysis across the MDD, we did not observe any consistent relationships between the best-performing simulation length and these features (see SI Fig. S6). Nevertheless, we note that more subtle dependencies may emerge when considering specific protein families, enzymatic classes, or systems involving pronounced allosteric motions, which remain important directions for future investigation.

Another important finding of this study is that even a rather generic MD protocol and a relatively small number of complexes can be used with success to achieve predictive accuracy on par with or better than highly complex models based on neural networks, with a much larger number of parameters. It is therefore expected that increasing the number of short MD simulations will further improve prediction performance. At scale, this conclusion brings hope to the inclusion of short MD simulations into protocols concerning diverse chemical library screening and hit prioritization and is also consistent with some previous studies done on single targets.[10,14,15]

The choice of traditional machine learning methods over deep learning was motivated by the tabular, physically interpretable nature of the descriptor set, as well as the need for transparency and robustness in the presence of correlated interaction features. While deep learning models have demonstrated strong performance in affinity prediction, the performance gap remains limited for engineered descriptors, and classical approaches provide a strong, interpretable baseline. However, the MDD provides MD trajectories that can be used to construct more sophisticated representations, including graph-based or sequence-aware encodings that are better suited for deep learning architectures for further comparison and benchmarking.

Comparisons between models trained on static crystallographic descriptors and those derived from MD trajectories reveal that the two approaches consider different descriptors to be most relevant. Interestingly, we note a number of time series derived descriptors with significantly better correlations compared to their static counterparts (Fig. 6). Their summarized influence, however, did transfer only slightly to improved affinity prediction performance. Given that extensive cross-correlation filtering was performed, this would suggest that non-linear correlations may decrease the overall performance. Ablation analysis and SHAP studies (see SI Fig. S3) further confirm these findings, showing that the two models employ different types of descriptors. In the case of more challenging target splits, an advantage of MD-derived models can also be observed, highlighting potential generalization advantages of the time series descriptors.

In both the screening and pose selection tasks, MD-based models showed better performance than models based solely on static crystallographic structures. However, the higher standard deviation in the results of the MD models suggests a greater sensitivity to the selection of the initial conformation, which may affect the reproducibility and stability of the prediction. This phenomenon may be due to the greater complexity of dynamic representations and should be further investigated, taking into account different classes of molecular targets and simulation conditions. Importantly, while the descriptors used in this study are derived from molecular dynamics trajectories, they represent statistical summaries of interaction properties sampled over time, rather than an explicit encoding of discrete dynamic events. The time series descriptors (ts_descriptors) capture ensemble-level tendencies, such as the persistence and variability of classical protein–ligand features, such as van der Waals contacts, hydrophobic interactions, hydrogen bonding, *etc.*, aggregated across the simulation window. Our analysis shows that the model's performance emerges from multivariate combinations of interaction features rather than individual interaction terms.

Consequently, the presented representation should be viewed as an approximation of interaction dynamics, designed with physical interpretability, robustness, and scalability for machine-learning applications. Capturing true dynamic processes, such as interaction lifetimes, state transitions, or allosteric shifts, would require alternative time-resolved or event-based representations, which remain an important direction for future work. Within this context, the MDD and the descriptor framework introduced here provide a reproducible and extensible baseline upon which more mechanistically explicit modeling approaches can be developed and benchmarked.

In summary, we demonstrate that short MD simulations combined with time-series descriptor representations and classical machine learning models can achieve predictive performance on par with, and in some cases exceeding, models based on static structures, while offering improved generalization and scalability. Although the developed MDD remains limited in size relative to static structures, our findings already show the strong potential of MD aware models and lay the foundation for further development of larger and more diverse protein–ligand MD collections.

Author contributions

JP was responsible for methodology design, software development, investigation and data curation. AY contributed to data interpretation and investigation. PS led the conceptualization of the study, methodology formulation, funding acquisition, and data curation and drafted the original manuscript. All authors have approved the final manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

Descriptor generation scripts are available from Github: https://github.com/JPoziemski/md_for_affinity_prediction and Zenodo: https://doi.org/10.5281/zenodo.18805105. Trajectories of molecular dynamics are deposited at Zenodo: https://doi.org/10.5281/zenodo.18805105. The PDBBind 2020 R1 dataset was downloaded from: https://www.pdbbind-plus.org.cn/. The DUD-E dataset was downloaded from: https://dude.docking.org.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5dd00452g.

Acknowledgements

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

References

1 V. Kairys, L. Baranauskiene, M. Kazlauskiene, D. Matulis and E. Kazlauskas, Binding affinity in drug design: experimental and computational techniques, *Expert Opin. Drug Discovery*, 2019, **14**, 755–768.
2 D. L. Mobley and M. K. Gilson, Predicting Binding Free Energies: Frontiers and Benchmarks, *Annu. Rev. Biophys.*, 2017, **46**, 531–558.
3 C. D. Parks, Z. Gaieb, M. Chiu, H. Yang, C. Shao, W. P. Walters, *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 P. J. Ballester and J. B. O. Mitchell, A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking, *Bioinformatics*, 2010, **26**, 1169–1175.
5 M. M. Stepniewska-Dziubinska, P. Zielenkiewicz and P. Siedlecki, Development and evaluation of a deep learning model for protein-ligand binding affinity prediction, *Bioinformatics*, 2018, **34**(21), 3666–3674.
6 Z. Wang, L. Zheng, Y. Liu, Y. Qu, Y.-Q. Li, M. Zhao, *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 M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, *et al.*, Comparative Assessment of Scoring Functions: The CASF-2016 Update, *J. Chem. Inf. Model.*, 2019, **59**, 895–913.
8 C. Shen, Y. Hu, Z. Wang, X. Zhang, H. Zhong, G. Wang, *et al.*, Can machine learning consistently improve the scoring power of classical scoring functions? Insights into the role of machine learning in scoring functions, *Briefings Bioinf.*, 2021, **22**, 497–514.
9 A. Ganesan, M. L. Coote and K. Barakat, Molecular dynamics-driven drug discovery: leaping forward with confidence, *Drug Discovery Today*, 2017, **22**, 249–269.
10 S. Gu, C. Shen, J. Yu, H. Zhao, H. Liu, L. Liu, *et al.*, Can molecular dynamics simulations improve predictions of protein-ligand binding affinity with machine learning?, *Briefings Bioinf.*, 2023, **24**(2), DOI: 10.1093/bib/bbad008.
11 D. Gioia, M. Bertazzo, M. Recanatini, M. Masetti and A. Cavalli, Dynamic Docking: A Paradigm Shift in Computational Drug Discovery, *Molecules*, 2017, **22**(11), DOI: 10.3390/molecules22112029.
12 P. Śledź and A. Caflisch, Protein structure-based drug design: from docking to molecular dynamics, *Curr. Opin. Struct. Biol.*, 2018, **48**, 93–102.
13 H. Guterres and W. Im, Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations, *J. Chem. Inf. Model.*, 2020, **60**, 2189–2198.
14 J. Ash and D. Fourches, Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories, *J. Chem. Inf. Model.*, 2017, **57**, 1286–1299.
15 S. Jamal, A. Grover and S. Grover, Machine Learning From Molecular Dynamics Trajectories to Predict Caspase-8 Inhibitors Against Alzheimer's Disease, *Front. Pharmacol*, 2019, **10**, 780.
16 P. P. Kyaw Zin, A. Borrel and D. Fourches, Benchmarking 2D/3D/MD-QSAR Models for Imatinib Derivatives: How Far Can We Predict?, *J. Chem. Inf. Model.*, 2020, **60**, 3342–3360.
17 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX*, 2015, **1–2**, 19–25.
18 N. Altman and M. Krzywinski, The curse(s) of dimensionality, *Nat. Methods*, 2018, **15**, 399–400.
19 RDKit, *Open-source cheminformatics*, Available from: https://www.rdkit.org.
20 N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, MDAnalysis: a toolkit for the analysis of molecular dynamics simulations, *J. Comput. Chem.*, 2011, **32**, 2319–2327.
21 C. Bouysset and S. Fiorucci, ProLIF: a library to encode molecular interactions as fingerprints, *J. Cheminform.*, 2021, **13**, 72.
22 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, *et al.*, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nat. Methods*, 2020, **17**, 261–272.
23 Overview on extracted features — tsfresh 0.20.2.post0.dev4+g3da2360 documentation, 2024, Available from: https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html.
24 B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing and Z. Wu, *Deep Learning for the Life Sciences*. O'Reilly Media, 2019.
25 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, *et al.*, Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

26 T. Chen and C. Guestrin, XGBoost A Scalable Tree Boosting System, *arXiv*, 2016, preprint, DOI: 10.48550/arXiv.1603.02754.
27 G. Menardi and N. Torelli, Training and assessing classification rules with imbalanced data, *Data Min Knowl Discov.*, 2014, **28**, 92–122.
28 M. Wójcikowski, M. Kukiełka, M. M. Stepniewska-Dziubinska and P. Siedlecki, Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions, *Bioinformatics*, 2019, **35**, 1334–1341.
29 J. D. Durrant and J. A. McCammon, NNScore 2.0: a neural-network receptor-ligand scoring function, *J. Chem. Inf. Model.*, 2011, **51**, 2897–2903.
30 O. Trott and A. J. Olson, 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.
31 Z. Cang, L. Mu and G.-W. Wei, Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening, *PLoS Comput. Biol.*, 2018, **14**, e1005929.
32 S. Zhang, Y. Jin, T. Liu, Q. Wang, Z. Zhang, S. Zhao, *et al.*, SS-GNN: A Simple-Structured Graph Neural Network for Affinity Prediction, *ACS Omega*, 2023, **8**, 22496–22507.
33 A. C. J. Mohamed, M. A. H. Newton, J. Rahman, A. J. Mohamed Abdul Cader and A. Sattar, Ensembling methods for protein-ligand binding affinity prediction, *Sci. Rep.*, 2024, **14**, 24447.
34 J. Wee and K. Xia, Ollivier Persistent Ricci Curvature-Based Machine Learning for the Protein-Ligand Binding Affinity Prediction, *J. Chem. Inf. Model.*, 2021, **61**, 1617–1626.
35 X. Liu, H. Feng, J. Wu and K. Xia, Dowker complex based machine learning (DCML) models for protein-ligand binding affinity prediction, *PLoS Comput. Biol.*, 2022, **18**, e1009943.
36 Z. Jin, T. Wu, T. Chen, D. Pan, X. Wang, J. Xie, *et al.*, CAPLA: improved prediction of protein-ligand binding affinity by a deep learning approach based on a cross-attention mechanism, *Bioinformatics*, 2023, **39**(2), DOI: 10.1093/bioinformatics/btad049.
37 X. Zhang, H. Gao, H. Wang, Z. Chen, Z. Zhang, X. Chen, *et al.*, PLANET: A Multi-objective Graph Neural Network Model for Protein-Ligand Binding Affinity Prediction, *J. Chem. Inf. Model.*, 2024, **64**(7), DOI: 10.1021/acs.jcim.3c00253.
38 J. Jiménez, M. Škalič, G. Martínez-Rosell and G. De Fabritiis, KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks, *J. Chem. Inf. Model.*, 2018, **58**, 287–296.
39 L. Zheng, J. Fan and Y. Mu, OnionNet: a multiple-layer intermolecular contact based convolutional neural network for protein-ligand binding affinity prediction, *arXiv*, 2019, preprint, arXiv:1906.02418, DOI: 10.48550/arXiv.1906.02418.
40 C. Shen, X. Zhang, C.-Y. Hsieh, Y. Deng, D. Wang, L. Xu, *et al.*, A generalized protein-ligand scoring framework with balanced scoring, docking, ranking and screening powers, *Chem. Sci.*, 2023, **14**, 8129–8146.
41 S. Chatterjee, A New Coefficient of Correlation, *J. Am. Stat. Assoc.*, 2021, **116**, 2009–2022.
42 M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking, *J. Med. Chem.*, 2012, **55**, 6582–6594.
43 S. M. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in *Advances in Neural Information Processing Systems*, ed. Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R. and Vishwanathan S., et al., Curran Associates, Inc., 2017, Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.
44 J. Sieg, F. Flachsenberg and M. Rarey, In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening, *J. Chem. Inf. Model.*, 2019, **59**, 947–961.
45 M. Volkov, J.-A. Turk, N. Drizard, N. Martin, B. Hoffmann, Y. Gaston-Mathé, *et al.*, On the Frustration to Predict Binding Affinities from Protein-Ligand Structures with Deep Neural Networks, *J. Med. Chem.*, 2022, **65**, 7946–7958.
46 P.-Y. Libouban, S. Aci-Sèche, J. C. Gómez-Tamayo, G. Tresadern and P. Bonnet, The Impact of Data on Structure-Based Binding Affinity Predictions Using Deep Neural Networks, *Int. J. Mol. Sci.*, 2023, **24**(22), DOI: 10.3390/ijms242216120.

