Supplementary material

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

Materials

Filtering Procedure

For the MDD dataset, consider only:

- targets not exceeding 400 AA
- targets with no missing AA, other than N-term or C-term AA
- binding site AA belonging to a single chain
- binding site with no missing AA
- binding site with no ions
- ligands not containing metal atoms, targets that are not metalloproteinase
- ligands that are not peptides
- ligands with unambiguous, experimentally determined affinity values (pKi, pKd, pIC50)

To maintain diversity, a cap was set at 18 protein complexes per target, identified by a common UniProtID. Complexes for which the MD simulation procedure (see below) returned errors were discarded.

MD simulation preparation

All MDD protein-ligand complexes were prepared following a standardized protocol. Missing atoms in the protein structures were added using the PDBFixer tool [1]. Protein targets were parameterized using the AMBER99SB-ILDN force field, while ligand parameterization was conducted with the ANTECHAMBER module within the ACPYPE tool [2]. For the ligands, partial charges were derived to match the quantum-mechanically generated electrostatic potential via the Restrained Electrostatic Potential (RESP) method [3], 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.

MDD assessment

Targets

Roughly around ⅔ of the MDD targets are enzymes with an assigned EC number, with hydrolases and transferases composing almost 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.

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

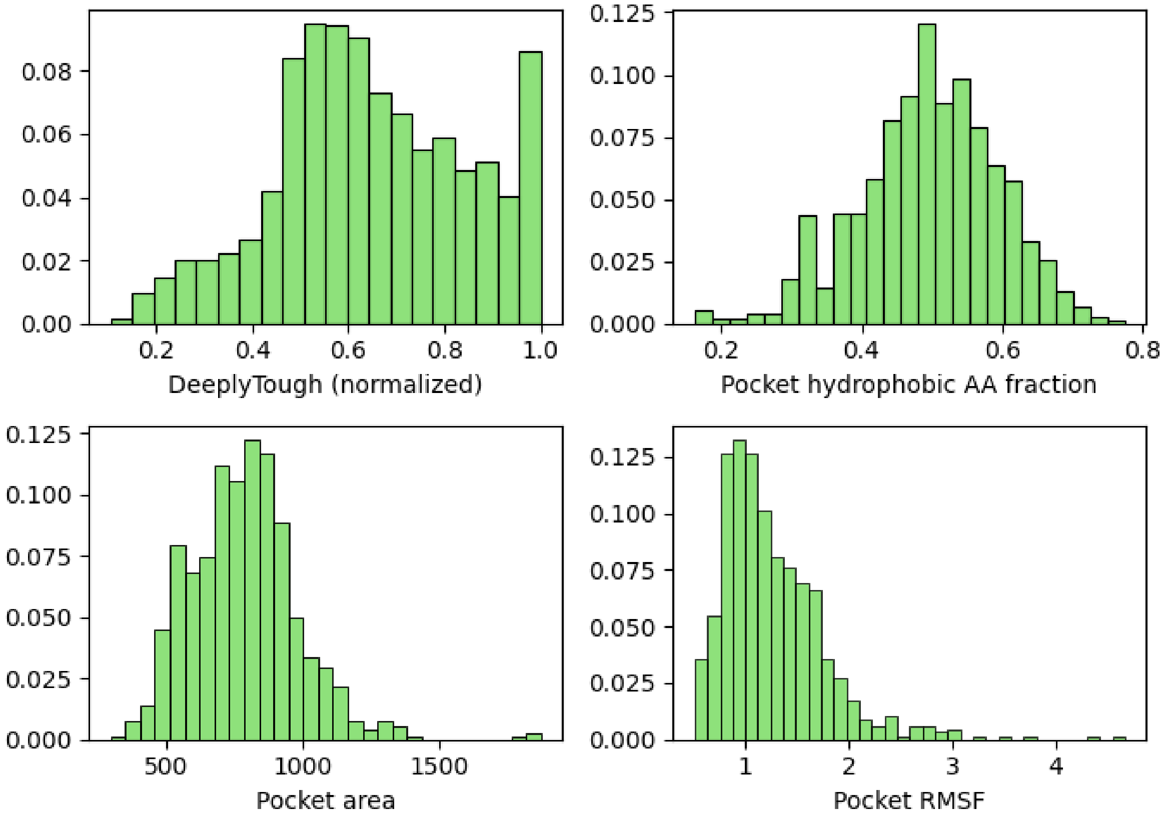

Figure S1: Comparison of binding site properties of MDD targets. Comparison with respect to structural similarity and ligand preference similarity (DeeplyTough), hydrophobic residues, size (area), and mobility (RMSF).

Ligand diversity

The affinity range of ligands in the MDD dataset is described roughly by a normal distribution (Figure 3, A) when the logarithm scale is used. 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 mean Tanimoto distance measured with the ECFP4 (1024 bits) fingerprint between all ligands in the dataset is 0.11 showing a low overall structural similarity of the chemical space. Similarity of ligands within their targets is also low: 0.30.

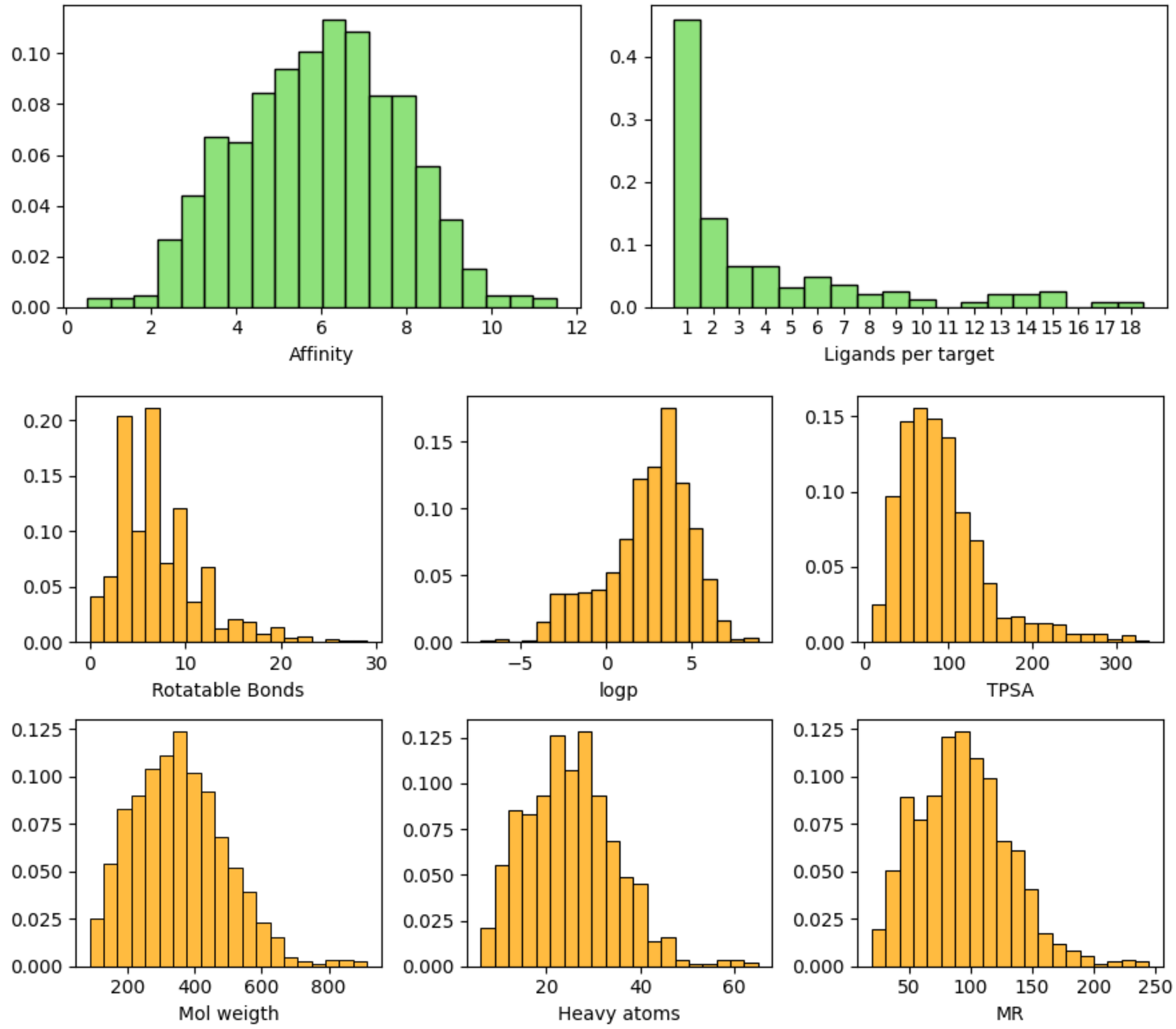

Figure S2: Ligand features distribution in the MDD dataset. Ligand per target shows the fraction of targets with a given number ligand complexes. Around half of the targets in MDD have a single ligand, 71 targets from 2 to 5 ligands, 31 from 5 to 10 and 22 with more than 10 ligands.

List of descriptors

Descriptors serve to represent the protein ligand complex. This work employs a variety of descriptors, categorized into four main groups: ligand descriptors, pocket (binding site) descriptors, interaction descriptors, and motion descriptors. Each complex from the MDD has two complementary representations; one derived from the crystallographic data and one from MD simulation.

Ligand property descriptors are features which do not change during a MD simulation. These are mainly ligand standard physicochemical features, calculated with RDKit v.2022.03.5 [4], such as the number of nitrogen atoms or the number of aromatic rings that remain constant throughout the analysis. All other descriptors (ligand, pocket, interaction and motion) can change their value during an MD simulation. Ligand geometric descriptors capture mainly spatial features such as eccentricity, radius of gyration, area and volume. To take into account ligand topology, ECFP4 was also added to the representation. While ligand geometry descriptors focus on the quantitative features such as distances and angles, ECFP4 describes the qualitative aspects of invariant properties such as connectivity and compactness. Unlike ligand property descriptors, pocket property descriptors can change during MD simulation. This is due to the binding pocket's definition being dependent on the ligand's position. As the ligand's position shifts during the MD simulation, the composition of the active site may also vary. The binding pocket is defined as comprising amino acids that have at least one atom within 6 Ångström (Å) proximity to at least one atom of the ligand. Pocket geometric descriptors, define two thresholds (3.5A and 5A) for calculating volume and area. ConvexHull function from SciPy package with pocket heavy atom positions was used to obtain the approximated values for both ligand and pocket geometric descriptors.

The interaction descriptors, describing protein-protein and protein-ligand contacts, are calculated with ProLIF [5] default settings. Motion descriptors can be calculated only from MD simulations and not from crystallographic data. These descriptors are specific to changes observed during the simulation, such as retaining or losing interactions, or changes to the RMSD and RMSF values of ligand and binding site pocket.

Table S1. Ligand property descriptors

no.	Name	Description	Calculated with
1	ligand_mol_weigth	Molecular weight	RDKit
2	ligand_logp	Wildman-Crippen logP	RDKit
3	ligand_hba	Number of hydrogen bond acceptors	RDKit

4	ligand_hbd	Number of hydrogen bond donors	RDKit
5	ligand_tpsa	Topological polar surface area	RDKit
6	ligand_mr	Molar refractivity	RDKit
7	ligand_arom_rings	Number of aromatic rings	RDKit
8	ligand_aliphatic_rings	Number of aliphatic rings	RDKit
9	ligand_rot_bounds	Number of rotatable bonds	RDKit
10	ligand_single_bonds	Number of single bonds	RDKit
11	ligand_double_bonds	Number of double bonds	RDKit
12	ligand_aromatic_bonds	Number of aromatic bonds	RDKit
13	ligand_N	Number of Nitrogen atoms	RDKit
14	ligand_C	Number of Carbon atoms	RDKit
15	ligand_O	Number of Oxygen atoms	RDKit
16	ligand_S	Number of Sulfur atoms	RDKit
17	ligand_H	Number of Hydrogen atoms	RDKit
18	ligand_P	Number of phosphorus atoms	RDKit
19	ligand_Halogen	Number of Halogen atoms	RDKit

Table S2. Ligand geometric descriptors

no.	Name	Description	Calculated with
1	ligand_area	Area estimated by convex hull based on ligand heavy atoms positions	scipy.spatial.ConvexHull
2	ligand_volume	Volume estimated by convex hull based on heavy atoms positions	scipy.spatial.ConvexHull
3	ligand_pmi1	First (smallest) principal moment of inertia	RDKit
4	ligand_pmi2	Second principal moment of inertia	RDKit
5	ligand_pmi3	Third (largest) principal moment of inertia	RDKit

6	ligand_rog	Radius of gyration	RDKit
7	ligand_pbf	Plane of best fit	RDKit
8	ligand_eccentricity	Molecular eccentricity	RDKit
9	ligand_asphericity	Molecular asphericity	RDKit

Table S3. Pocket property descriptors

No.	Name	Description	Calculated with
1	pocket_arom_rings	Number of aromatic rings	RDKit
2	pocket_aliphatic_rings	Number of aliphatic rings	RDKit
3	pocket_rot_bonds	Number of rotatable bonds	RDKit
4	pocket_single_bonds	Number of single bonds	RDKit
5	pocket_double_bonds	Number of double bonds	RDKit
6	pocket_aromatic_bonds	Number of aromatic bonds	RDKit
7	pocket_N	Number of Nitrogen atoms	RDKit
8	pocket_C	Number of Carbon atoms	RDKit
9	pocket_H	Number of Hydrogen atoms	RDKit
10	pocket_O	Number of Oxygen atoms	RDKit
11	pocket_S	Number of Sulfur atoms	RDKit
12	pocket_aliphatic	Number of aliphatic amino acids: ALA, ILE, LEU, PRO, VAL	MDAnalysis
13	pocket_hydrophobic	Number of hydrophobic amino acids: ALA, ILE, LEU, MET, PHE, VAL, PRO, GLY	MDAnalysis
14	pocket_charged	Number of charged amino acids: ARG, LYS, ASP, GLU	MDAnalysis
15	pocket_aromatic	Number of aromatic amino acids: PHE, TRP, TYR	MDAnalysis
16	pocket_polar	Number of polar amino acids: GLN, ASN, HIS, SER, THR, TYR, CYS	MDAnalysis
17	pocket_hba	Number of hydrogen bond acceptors	RDKit

18	pocket_hbd	Number of hydrogen bond donors	RDKit

Table S4. Pocket geometric descriptors

No	Name	Descriptor	Calculated with
1	pocket_3_5_volume	Volume estimated by convex hull based on pocket heavy atoms positions. Pocket defined as <=3.5A distance of any ligand heavy atom.	scipy.spatial.ConvexHull
2	pocket_5_volume	Volume estimated by convex hull based on pocket heavy atoms positions. Pocket defined as <=5 A distance of any ligand heavy atom.	scipy.spatial.ConvexHull
3	pocket_3_5_area	Area estimated by convex hull based on pocket heavy atoms positions. Pocket defined as <=3.5 A distance of any ligand heavy atom	scipy.spatial.ConvexHull
4	pocket_5_area	Area estimated by convex hull based on pocket heavy atoms positions. Pocket defined as <=5 A distance of any ligand heavy atom	scipy.spatial.ConvexHull
5	pocket_pocket_contact_all	Number of pocket internal contacts defined as < 4,5A between pocket heavy atoms.	MDAnalysis
6	ligand_pocket_contact_all	Number of ligand-pocket contacts defined as < 4,5A between pocket heavy atoms.	MDAnalysis

Table S5. Interaction descriptors

No	Name	Description	Calculated with
1	vdWContact	documentation link	ProLIF
2	Hydrophobic	documentation link	ProLIF
3	HBAcceptor	documentation link	ProLIF
4	Anionic	documentation link	ProLIF

5	CationPi	documentation link	ProLIF
6	EdgeToFace	documentation link	ProLIF
7	FaceToFace	documentation link	ProLIF
8	PiCation	documentation link	ProLIF
9	PiStacking	documentation link	ProLIF
10	XBDonor	documentation link	ProLIF
11	HBDonor	documentation link	ProLIF

Table S6. Motion descriptors

Pocket defined as all amino acids that any atom was at distance <= 3.5 from any ligand heavy atoms. Only heavy atoms are taken into account.

No	Name	Description	Calculated with
1	contact_pocket_old	fraction of pocket internal contacts preserved between processed frame and first frame (reference).	MDAnalysis
2	contact_pocket_new	fraction of pocket internal contacts present in the processed frame compared to last frame (reference).	MDAnalysis
3	contact_ligand_pocket_old	fraction of pocket-ligand contacts preserved between processed frame and first frame (reference)	MDAnalysis
4	contact_ligand_pocket_new	fraction of pocket-ligand contacts present in the processed frame compared to last frame (reference)	MDAnalysis
5	RMSD_pocket	Residual mean square deviation of pocket atoms compared to previous frame	MDAnalysis
6	RMSD_ligand	Residual mean square deviation of ligand atoms compared to previous frame	MDAnalysis
7	RMSD_ca	Residual mean square deviation of pocket carbon alpha atoms compared to	MDAnalysis

		previous frame	

Model parameters

Descriptor model (XGB)

- `n_estimators=512,`
- `eta=0.05,`
- `max_depth=10,`
- `min_child_weight=0.7`
- `colsample_bytree=0.9`
- `colsample_bylevel=0.9`
- `colsample_bynode=0.9,`
- `alpha=0.2`

Parameter names consistent with the `xgboost` python package.

Random Forest

Model parameters where obtained with the use of parameter matrix:

- `min_samples_leaf = [2, 3,4, 5, 6]`
- `min_samples_split = [2, 3,4, 5,6, 7, 8]`

Number of trees was set to 268.

All other parameters were set to default.

Figure S3

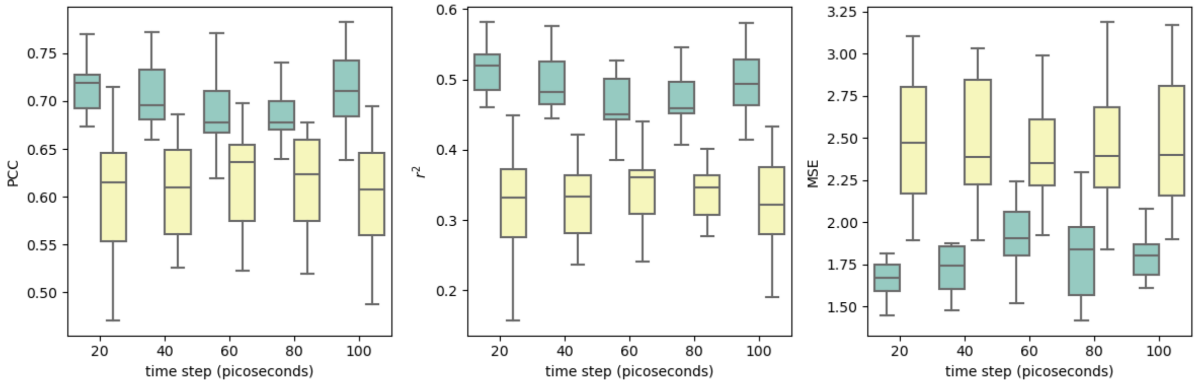

Figure S3: influence of sampling frequency on models performance in two splits.

Results based on 20ns MDs.

SHAP analysis

Table S7 - 20 most influential descriptors and ts_descriptors.

Time series descriptors (ts_descriptors)	Model count
pocket_rot_bonds__abs_energy	20
pocket_O__cwt_coefficients__coeff_9__w_10__widths_(2, 5, 10, 20)	20
pocket_hydrophobic__c3__lag_1	20
pocket_C__root_mean_square	20
pocket_3_5_volume__c3__lag_3	20
pocket_aromatic__linear_trend__attr_"intercept"	20
ligand_pmi1__c3__lag_3	20
ligand_logp	20
ligand_asphericity__mean_n_absolute_max__number_of_maxima_7	20
contact_ligand_pocket_old__standard_deviation	20
contact_pocket_pocket_new__sum_values	20
contact_pocket_all__benford_correlation	20
Hydrophobic__benford_correlation	20
pocket_5_area__fft_aggregated__aggtype_"skew"	18
pocket_arom_rings__benford_correlation	18
pocket_double_bonds__cwt_coefficients__coeff_5__w_5__widths_(2, 5, 10, 20)	17
contact_ligand_pocket_all__cwt_coefficients__coeff_9__w_10__widths_(2, 5, 10, 20)	12

ligand_pmi2__ar_coefficient__coeff_0__k_10	12
ligand_eccentricity__eccentricity__sum_values	10
RMSD_ligand__fft_aggregated__aggtype_"kurtosis"	9

Static descriptors	Model count
pocket_aliphatic	20
pocket_5_area	20
pocket_hydrophobic	20
pocket_aromatic	20
pocket_3_5_area	20
pocket_3_5_volume	20
pocket_5_volume	20
ligand_logp	20
ligand_pbf	20
ligand_tpsa	20
contact_pocket_all	20
ligand_mol_weigth	19
ligand_mr	18
contact_ligand_pocket_all	16
ligand_pmi1	16
ligand_eccentricity	16
Hydrophobic	16
ligand_asphericity	15
ligand_Halogen	13
pocket_C	8

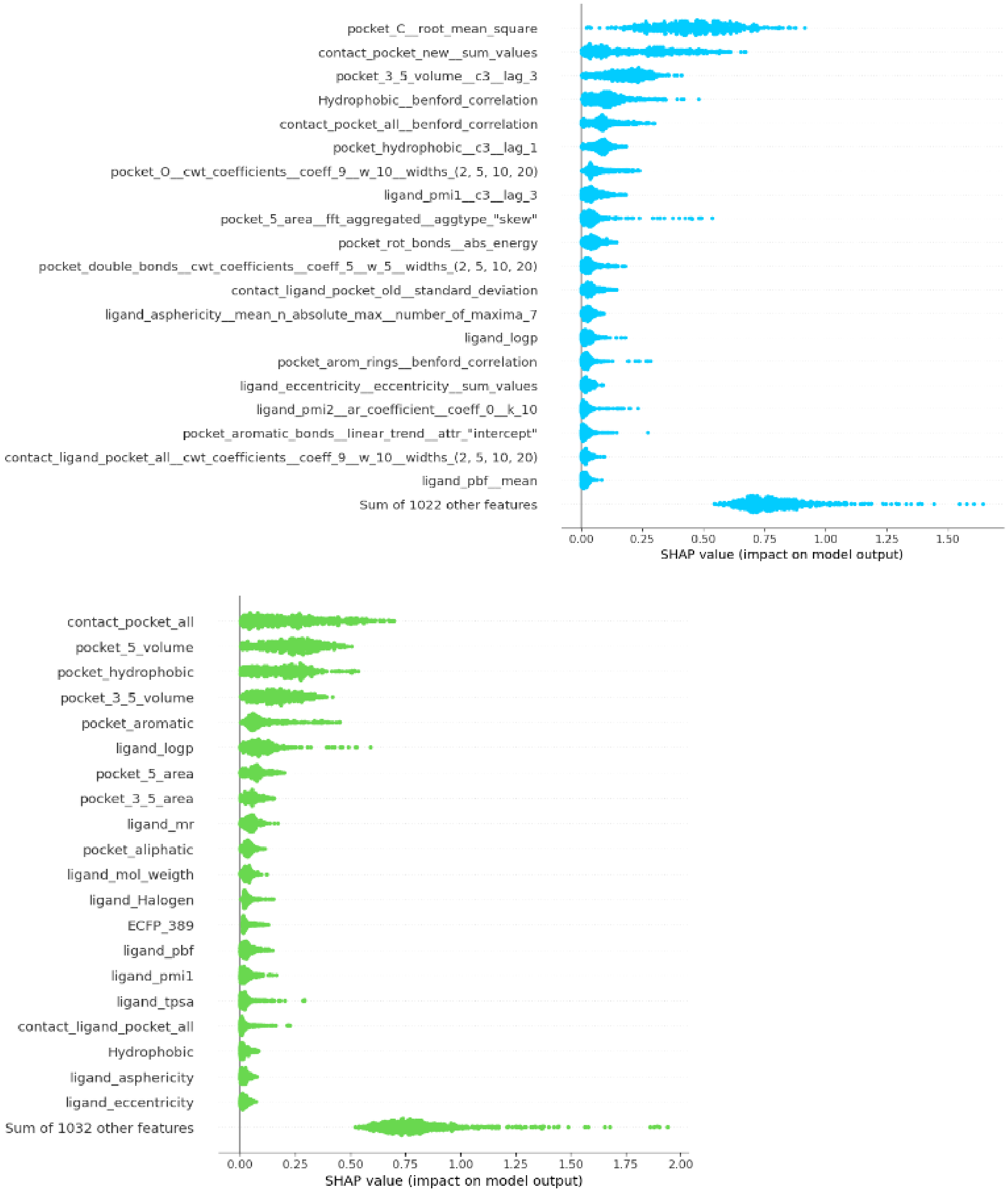

Figure S4: Example SHAP analysis for a single RFmodel. Blue panel - 20 most important time series descriptors (ts_descriptors). Green panel - 20 most important static descriptors.Only half of pocket descriptors and ts_descriptors are counterparts. However, two most important contact descriptors (pocket internal contacts and pocket-ligand contacts) are similarly important for both models. The same is true for the sole hydrophobic descriptor, the only interaction type present in the top 20.

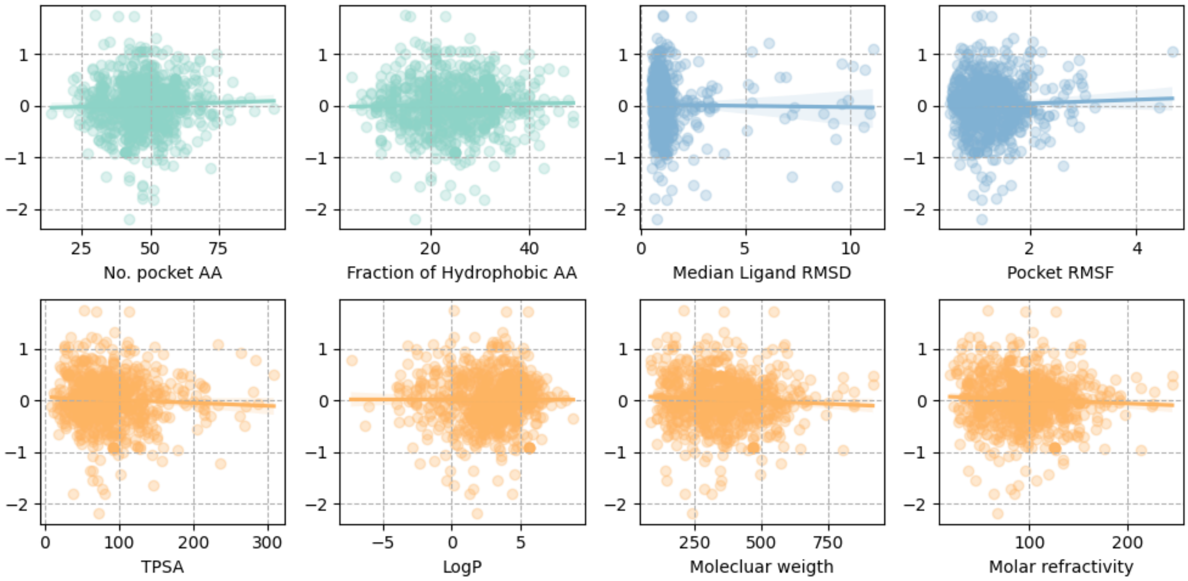

Figure S5. Difference in affinity prediction absolute error compared with selected ligand (orange) pocket (green) motion (blue) complex features. The values on the Y axis represent the difference in absolute errors. Points with positive values represent complexes for which static representation was better, negative values on the Y axis are complexes for which the MD-derived model performed better. There is no clear correlation between the mobility of the binding pocket (RMSF) nor the ligand movements (RMSD), and the performance of both models. The obtained result is different from that presented by [6], who postulated that MD augmented models should perform better when dealing with more flexible complexes. However, the points on the Y axis take on a variety of values, thus both models can perform substantially differently with individual complexes.

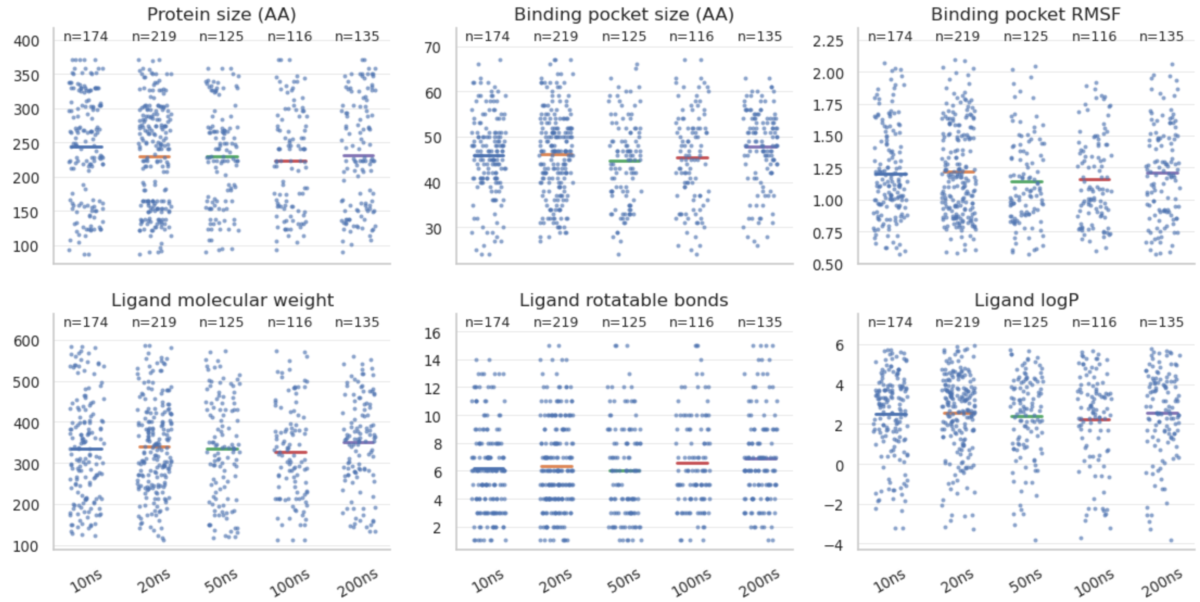

Figure S6. Relationship between simulation length and selected macroscopic protein-ligand features in the context of model performance. The swarmplots show individual data points with the mean indicated. Each point represents a single protein-ligand complex from the MDD dataset, for which the highest affinity prediction performance was achieved at a given simulation length. The results indicate that the optimal simulation length is not trivially determined by system size nor intrinsic flexibility. For clarity, relative performance gains between different simulation lengths are not shown in this figure; these can be compared in Figure 3 in the Simulation length section of the main text.

References

1. 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.

2. 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

3. 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.

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

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

6. 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