Journal of Computer-Aided Molecular Design (2024) 38:24
https://doi.org/10.1007/s10822-024-00564-2

MDFit: automated molecular simulations workflow enables high throughput assessment of ligands-protein dynamics

Alexander C. Brueckner[1] · Benjamin Shields[1] · Palani Kirubakaran[2] · Alexander Suponya[1] · Manoranjan Panda[1] · Shana L. Posy[1] · Stephen Johnson[1] · Sirish Kaushik Lakkaraju[1]

Received: 23 January 2024 / Accepted: 28 June 2024

Abstract
Molecular dynamics (MD) simulation is a powerful tool for characterizing ligand–protein conformational dynamics and offers significant advantages over docking and other rigid structure-based computational methods. However, setting up, running, and analyzing MD simulations continues to be a multi-step process making it cumbersome to assess a library of ligands in a protein binding pocket using MD. We present an automated workflow that streamlines setting up, running, and analyzing Desmond MD simulations for protein–ligand complexes using machine learning (ML) models. The workflow takes a library of pre-docked ligands and a prepared protein structure as input, sets up and runs MD with each protein–ligand complex, and generates simulation fingerprints for each ligand. Simulation fingerprints (SimFP) capture protein–ligand compatibility, including stability of different ligand-pocket interactions and other useful metrics that enable easy rank-ordering of the ligand library for pocket optimization. SimFPs from a ligand library are used to build & deploy ML models that predict binding assay outcomes and automatically infer important interactions. Unlike relative free-energy methods that are constrained to assess ligands with high chemical similarity, ML models based on SimFPs can accommodate diverse ligand sets. We present two case studies on how SimFP helps delineate structure–activity relationship (SAR) trends and explain potency differences across matched-molecular pairs of (1) cyclic peptides targeting PD-L1 and (2) small molecule inhibitors targeting CDK9.

Keywords molecular dynamics · machine learning · automated workflows · simulation fingerprints · pocket dynamics · MD · ML

Introduction

Structure-based drug design (SBDD) has become central to the drug discovery process and helped identify several marketed drugs available today [1]. Physics-based computational approaches that characterize protein–ligand interactions have significantly evolved [2] and benefited immensely from advances in hardware and algorithm optimizations [3]. Among the wide gamut of physics-based SBDD approaches, docking methods [4] continue to be among the most popular and have been used for a range of drug discovery processes including library screening [5] and ligand optimization [6]. Although their primary appeal lies in the ability to quickly predict the binding pose of a ligand in the protein pocket, it has been shown repeatedly that incorporating conformational dynamics of protein–ligand interactions is critical for driving the ligand optimization process [7].

Molecular dynamics (MD) simulations are an important tool for understanding the dynamics of binding pockets and optimizing ligands for drug discovery [8]. MD simulations can provide detailed information about the dynamic behavior of proteins and their interactions with ligands [9]. MD simulations reveal the stability of the complex and identify potential weaknesses or vulnerabilities that are useful in ligand optimization. MD simulations have been critical for delineating the relation between pocket dynamics and function of several classes of proteins including transmembrane receptors like ion channels [10], opioid receptors [11], viral capsids [12], sirtuins [13], and RAS [14] family proteins.

✉ Alexander C. Brueckner
brucknera15@gmail.com

✉ Sirish Kaushik Lakkaraju
kaushik.lakkaraju@bms.com

[1] Molecular Structure & Design, Bristol Myers Squibb, Princeton, NJ 08540, USA

[2] Biocon Bristol Myers Squibb R&D Centre, Bangalore 560099, Karnataka, India

Published online: 17 July 2024

These studies led to the development of selective activators or inhibitors [15–18] for these protein targets.

While there have been significant advances in high-performance computing infrastructure [19–22] and optimization of MD algorithms [23–25] to enable running MD with biological systems of increasing size [26–28] and complexity [29–31], the process of setting up, running, and analyzing data from MD simulations continues to be multi-step [32, 33] and cumbersome. This severely constrains the regular use of MD for compound prioritization in optimization campaigns typically run in industry. Moreover, several recent works have started adopting different strategies to dramatically increase chemical search space considered either via generative machine learning (ML) strategies [34] or through docking exercises involving extremely large libraries [35, 36] in screening and optimization cycles of discovery projects [37]. These studies have applied thermodynamic methods to enrich hit rates by accounting for dynamic protein–ligand interactions and conformational heterogeneity of the protein and ligand and the interplay with water [38, 39]. Given the limitations around the chemical similarity of compounds considered in a dataset for relative free energy calculations [40] and conformational sampling with thermodynamic approaches [41], incorporating 'regular' long-time-scale MD into assessing large libraries from generative ML or enumeration workflows will improve the accuracy of predictions and increase enrichment of hits from these workflows.

We present an automated workflow (MDFit) that streamlines setting up, running, and analyzing Desmond [25, 42] MD simulations of protein–ligand complexes using the OPLS4 [43] force field. The workflow takes a library of pre-docked ligands and a protein structure as input, sets up and runs MD with each of the protein–ligand complexes, and then analyzes MD trajectories of each of the ligands in the input dataset. Analysis of MD trajectories includes flexibility of the ligand in the pocket via root mean squared deviation (RMSD) compared to the starting pose, stability of different ligand-pocket interactions, and other useful metrics that help quantify the dynamics of protein pocket and the ligand library. These metrics are combined into simulation fingerprints (SimFPs) that enable easy rank-ordering of the dataset along any of these collected metrics. In addition, we demonstrate that SimFPs can be used as features in ML models for potency prediction and mechanistic interpretation. In contrast to static encodings like protein–ligand interaction fingerprints, SimFPs capture the dynamics of protein–ligand interaction and facilitate more accurate predictions. Unlike relative free energy perturbation calculations, SimFP-based ML models are less restrictive about the need for chemical similarity within a dataset and can accommodate much more comprehensive sampling of pocket ligand dynamics through longer time scale MD. While there have been some attempts in the past with automating MD, analyzing bulk phase behavior of ligands [44, 45] or analyzing protein- ligand interactions [33, 46–48], our workflow streamlines and integrates all these aspects towards enabling a potency prediction ML model that learns comprehensively about protein–ligand interactions in the presence of water from MD and explains interesting SAR trends that are otherwise missed from static structure-based methods.

We show applications of MDFit for assessing (a) cyclic peptides that target PD-L1 & (b) small molecule inhibitors targeting CDK9, both with therapeutical potential as anticancer agents [49–51]. Compound names (Pep-01 to Pep-60 for PD-L1 [51] and Cpd 01 [52] to Cpd 39) from the original publications are retained.

PD-L1 binds to PD-1 at an elongated β-sheet interface. Cyclic peptides with beta-strand geometry offer unique advantages for binding to this shallow and expansive orthosteric site. An overlay of Pep-01 bound to PD-L1 (PDB code 6PV9) and PD-1 bound to PD-L1 (PDB code 4ZQK) shows that Pep-01 binds to the β-sheet interface between PD-1 and PD-L1 (Fig. 1A). By mimicking the PD-1 secondary structure, Pep-01 packs against the PD-L1 surface with sufficient interaction energy to overcome the major costs of binding (Fig. 1B).

Previous studies have shown a strong correlation between peptide strain and their potency through docking of extensively sampled conformations of the peptides [53]. The extremely large number of rotatable dihedrals with these cyclic peptides makes relative free-energy perturbation methods for assessing potency and pocket dynamics untenable [54]. We apply MDFit to provide insights from protein-peptide dynamics that can clearly explain potency cliffs among matched-molecular pairs (MMPs). SimFPs enable easy identification of differences in the pocket and water-mediated interactions across MMPs that help build an understanding of the structure–activity relationship (SAR). In this study, SimFP features are also used for training an ML model to predict potency outcomes and infer which features are most important for activity. For the PD-L1 dataset, the top SimFP features identified by the ML model offer additional insights about MMPs and their potency cliffs that would have otherwise been easy to miss with static information such as docked poses.

Cyclin-dependent kinases (CDKs) are Ser/Thr kinases regulated by cyclins. Several CDK inhibitors have advanced to the clinic and have shown efficacy for multiple myeloma and other tumors [52]. We ran MDFit with a series of azabenzimidazole inhibitors [51] using a previously published co-crystal structure (Fig. 1C). Akin to the PD-L1 data set, top SimFP features identified from MDFit helped explain interesting SAR trends among MMPs that were otherwise not immediately apparent from their docked poses.

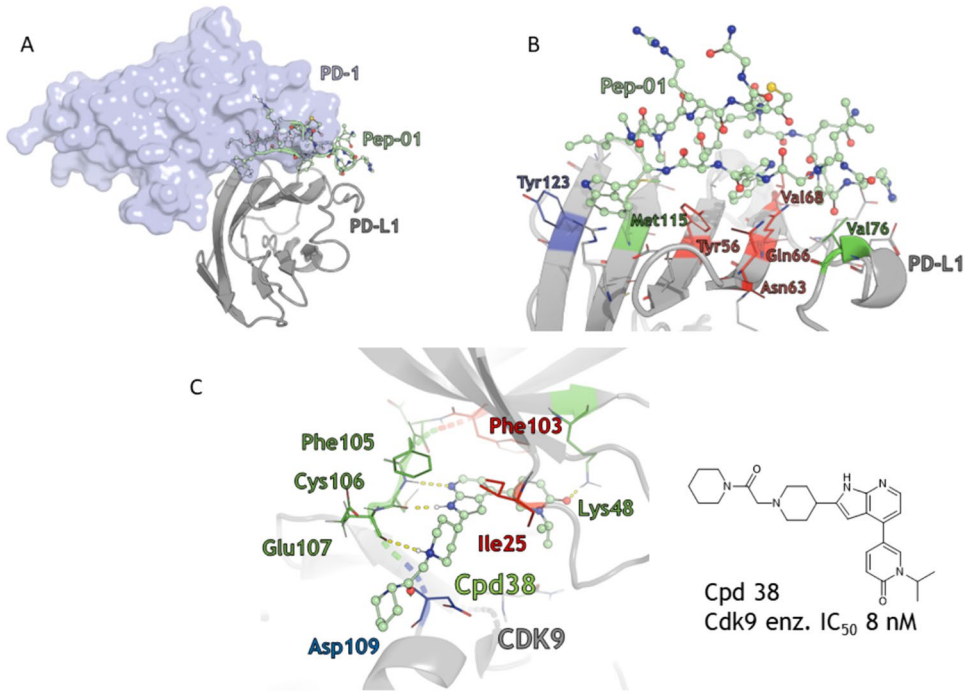


Fig. 1 **A** Overlay of 4ZQK and 6PV9 crystal structures showing Pep-01 binds to the β-sheet interface of PD-L1 to block PD-1 binding. **B** Peptide binding interface with PD-L1. All residues within 5 Å of Pep-01 are shown with critical residues determined by ML models (vide infra) shown in red (detrimental), green (beneficial), or blue (detrimental or beneficial). **C** Pocket interactions of Cpd 38 in CDK9 using PDB 7NWK. While docked poses were indistinguishable within the series, MDFit was useful in identifying the detrimental effect of pushing against Phe 103 & Ile 25 (vide infra)

Methods

The MDFit workflow (Fig. 2) automates the following process and the scripts are available for download from Github (https://github.com/brueckna2020/MDFit). The workflow requires the user to provide a protein model and a library of ligands as inputs. The protein structure needs to be fully prepared with missing side chains and loops added, protonation states of residues determined, hydrogen atoms added, and terminal residues capped. For the PD-L1 case study, protein from PDB 6PV9 [55] was used as the starting protein conformation. For the CDK9 case study, protein from PDB 7NWK was used as the starting protein conformation. Protonation states of protein residues were determined using PropKa [56] and

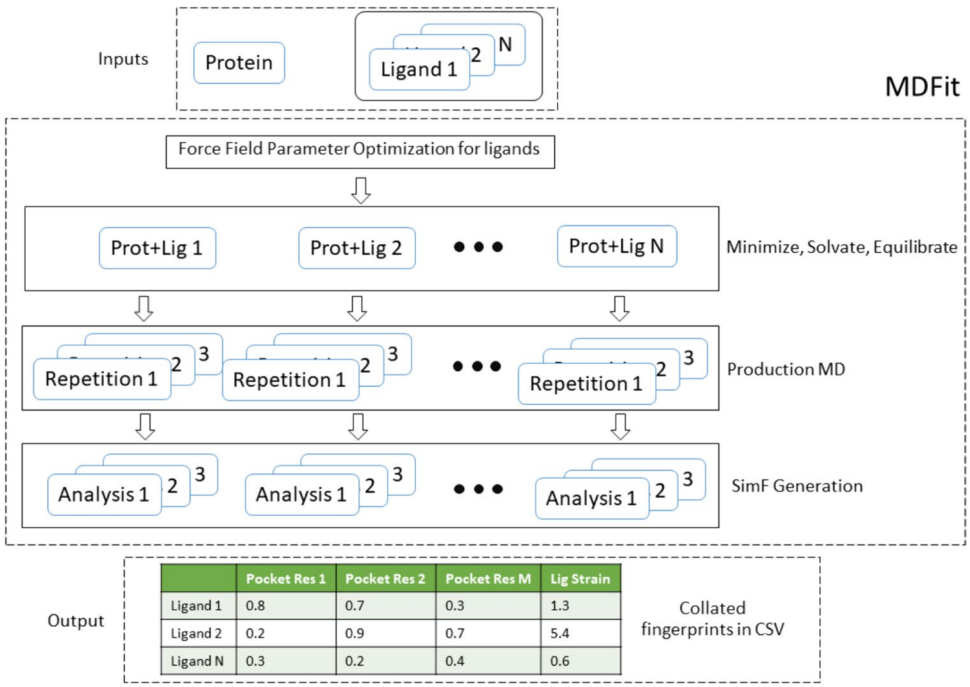


Fig. 2 MDFit workflow takes a library of ligands with reasonable starting poses in a protein pocket, runs MD, and generates collated SimFP for easy analysis of the stability of all ligands in the protein pocket across MD trajectories

the protein was prepared using the Protein Preparation Wizard module in Maestro (Schrodinger, LLC). Ligands in the input library need to have 3D conformations with reasonable poses when bound to the protein pocket. For the PD-L1 dataset, Pep-01 [57] and sixty of its analogs as described previously [57] were used. Previous studies have harnessed solution-state NMR and X-ray co-crystal structures of Pep-01 to accurately generate bound states of Pep-01 and its analogs [53, 57, 58]. Top poses from the docked conformer ensembles [53] were used as starting conformations for MDFit. For the CDK9 case study, thirty-nine azabenzimidazole inhibitors described previously [51] were used. Compounds were docked into the pocket using Glide [59] and poses similar to Compound 6 in the crystal structure were used as inputs for MDFit.

1) Force-field parameters: The workflow begins with a call to the FFBuilder tool from Schrodinger that evaluates all dihedrals in the input library, sets up QM calculations for dihedral scans, and optimizes missing or sub-optimal dihedral parameters using these QM scans. Optimized dihedral parameters are merged into the OPLS4 [43] 'main' force field supplied by the user. This optimized force field is subsequently used for MD and analysis.
2) Protein–ligand complexes: Each of the ligands in the input library is complexed with the protein which is put through an initial round of minimization using the MacroModel [60] module by Schrodinger. Powell-Reeves Conjugate Gradient (PCRG) minimization of the complex is run for a maximum of 500 steps with a convergence criterion of all gradient thresholds set to 0.3 kJ/mol.
3) Solvation: Minimized protein–ligand complexes are then inserted into an orthorhombic box with dimensions determined to set each edge of the box at 10 Å from the protein surface. The total charge of the protein and ligand is calculated and neutralizing ions Na^+ or Cl^- are placed randomly inside the box between the protein surface and the box edges. The remaining space is filled with water molecules.
4) Relaxation, Equilibration:
    a. Protein, ligand, and ion parameters are modeled using OPLS4 [43] while SPC [61] is used to model water. All simulations are run using the Desmond [25] engine. Both the case studies discussed below were run with Desmond from Schrodinger suite version 2022-2.
    b. Solvated protein–ligand systems are relaxed before the production MD simulations. Initially, the entire system is equilibrated for 100 ps using the NVT Brownian dynamics at T = 10 K, with a harmonic position restraint of force constant of 50 kcal/mol/Å^2 applied to all protein & ligand heavy atoms. At the same temperature and using the same restraints, the system is equilibrated for an additional 24 ps using a Berendsen [62] thermostat with pressure gradually dropping from 50 to 2 bars through NPT dynamics run.
5) Production: After equilibration, production MD simulations are run using NPT dynamics without positional restraints. By default, the workflow is set to run each protein–ligand solvated system in triplicate for a simulation time of 2 ns with a trajectory saving frequency set to 100 ps. Velocity seeds are randomized for each of the three MD runs. While the default settings stand at 2 ns for disk space considerations since MD trajectory files can be quite large, our calculations with PD-L1 & CDK9 data sets show that running simulations to 100 ns helps with convergence (see section on Simulation Length) and capturing interesting SAR trends. Therefore, for the PD-L1 & CDK9 datasets, each ligand–protein system was run for 3 replicates, each for 100 ns.
6) Analysis: Schrodinger's Simulation Event Analysis (SEA) scripts are used for assessing the production MD trajectories. The scripts collect a wide range of metrics (Supplementary Info, Table SI) that capture meaningful information and insights about ligand and pocket flexibility.
    a. Clusters from Trajectories: RMSD-based clustering analysis provides the top N cluster representations (default of 5) of the model system, revealing common structural motifs or states. The Desmond MD clustering algorithm calculates the RMSD similarity matrix for the given trajectory frames. By default, ligand atoms are used for RMSD calculations, and the matrix is computed based on these chosen atoms. Subsequently, the workflow clusters the trajectory frames using the RMSD matrix. An affinity propagation algorithm is employed for clustering, which is well-suited for identifying distinct conformational clusters. The output CMS files include information about cluster size, frame indices, and timestamps. These diverse conformations based on ligand RMSD are used for all analyses described with the PD-L1 dataset.
    b. Parched Trajectory: A trimmed MD trajectory is generated by retaining only the protein + ligand and closest N solvent molecules. By default, this is set to 100. Before parching, trajectories are aligned using the ligand atoms from the starting pose for reference.

c. Interactions: Protein–ligand interactions, water-mediated interactions, dihedral motions in ligands, and ion permeation are all recorded using event detection scripts that use pre-defined distance, angle, and dihedral cutoffs based on literature precedent [63–65]. The workflow extracts and tabulates all protein–ligand interactions and characterizes their stability as a percentage of the simulation time that each interaction was observed. For the PD-L1 dataset, along with the protein–ligand interactions, pre-calculated strain from the docked pose [53] is also added to the SimFP output for further analysis. Although all frames of the triplicate MD production runs can be included for this analysis (and is the default setting in the workflow), for both PD-L1 & CDK9 datasets, the first 10 ns (100 frames) of MD with each ligand were not considered for fingerprint generation.

While the automated part of the MDFit workflow stops with the generation of SimFPs, a predictive model can be readily trained to map SimFPs to experimental potency values. We emphasized the selection of simple, interpretable models that enable both the prediction of potency from SimFPs and the identification of important features. In this study, we investigated Linear, Ridge, Lasso, Random Forest, and Gradient Boosting Regression as implemented in scikit-learn [66] (see Supporting Information; Table S2, Figures S1–S5). Our workflow uses regression weights, impurity for tree-based models, and/or leave-one-feature-out cross-validation to estimate feature importance. Model prediction performance was investigated via nested leave-one-molecule-out cross-validation (LOMO-CV). SimFPs from triplicate runs were used as- is or averaged to arrive at an input feature matrix. The feature matrix was preprocessed by normalizing on the unit hypercube. The target IC_{50} values were transformed to pIC_{50} values and standardized to zero mean and unit variance. For each model type, hyperparameters were selected by minimizing the mean squared error using grid search LOO-CV. Feature importance was computed in CV folds and a final model was fit to the full data set for comparison.

Results and discussion

Simulation fingerprints (SimFPs) are a collection of interactions between a ligand and the protein target observed through MD simulations. The reported values are the average interaction frequency across a simulation. For example, a SimFP of 0.5 translates into a protein–ligand interaction occurring in 50% of the MD simulation frames. A SimFP value can be greater than 1.0 in cases where a ligand interacts with a protein residue through multiple points of contact (e.g., a bidentate interaction).

SimFPs can be used to rank-order or identify patterns across Matched Molecular Pairs (MMPs) for ligands with experimental readouts. Observed trends can be used to prioritize design ideas where the user gives preference to those that retain or enhance desired interactions. For larger data sets, SimFPs can be used as features to train ML models that can in turn be used to predict experimental readouts and assign feature importance. In addition, the critical SimFPs highlighted by the ML model can be used to further explain differences in observed readouts, such as potency. In this section, we discuss the utility of SimFPs in detail, focusing first on feature importance followed by handling edge cases.

SimFP feature identification

Machine learning methods can be used to identify specific peptide-protein interactions that contribute to the prediction of the desired endpoint from the full SimFP data set. For PD-L1, a Lasso regression model was built to predict the HTRF pIC_{50} values using SimFPs and strain energy [53] as features. While the model performance was modest (Fig. 3, right; LOMO-CV $Q^2 = 0.36$ and RMSE = 0.78), using the SimFPs as features provides interpretability lost in more complex modern ML models.

The top ten features (weights with the largest absolute value) of the PD-L1 data set are reported in Fig. 3, left. We note that along with interaction stability fingerprints that come from MDFit, pre-computed strain energies [53] were included as an additional feature of SimFP. Strain energy remains the standout feature, consistent with previous studies [53], while a water-mediated interaction with Asn63 was the most detrimental (negative weight) and a water-mediated interaction with Val76 was the most beneficial (positive weight) SimFP to potency. Based on the feature importance, peptide optimization should focus heavily on minimizing peptide strain followed by minimizing water-mediated interactions with Asn63 and maximizing water-mediated interactions with Val76. Select Match Molecular Pair (MMP) cases will be described herein using the feature selection to explain SAR.

MMPs with strain energy differences

Mutating position 2 from NMe-Ala in Pep-01 to NMe-Val in Pep-41 results in a significant drop in potency ($pIC_{50} = 8.1$ vs 6.0, respectively). Minor variations were observed for the top SimFP features, but a major increase in strain energy for Pep-41 explains the loss in potency (Fig. 4). While seemingly minor, the addition of a bulky sidechain distorted Pep-41's backbone conformation, increasing the strain by nearly 0.02 kcal/mol/heavy atom which is a remarkably high cost

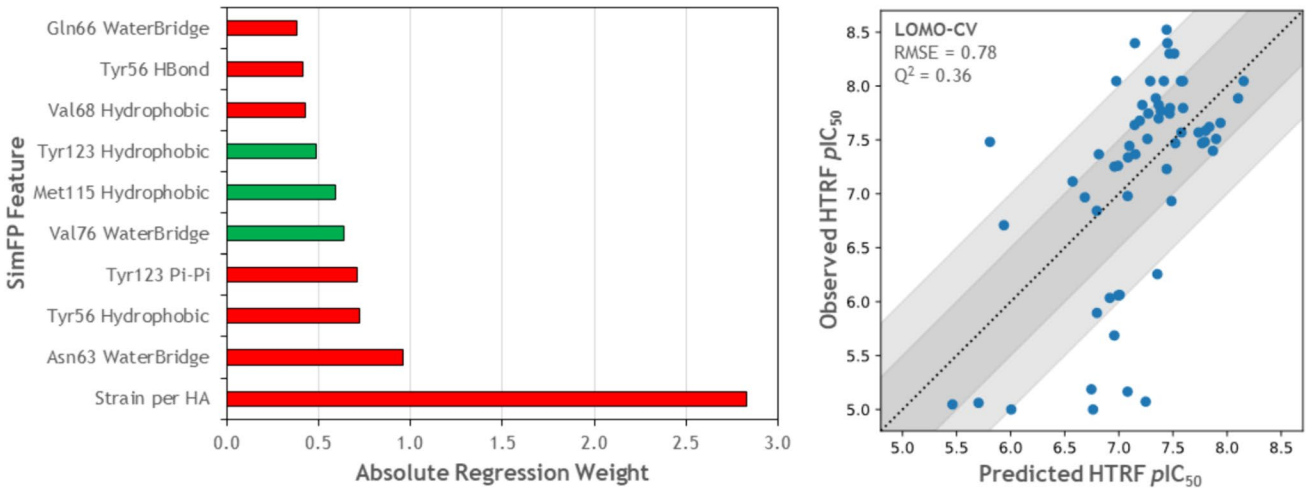
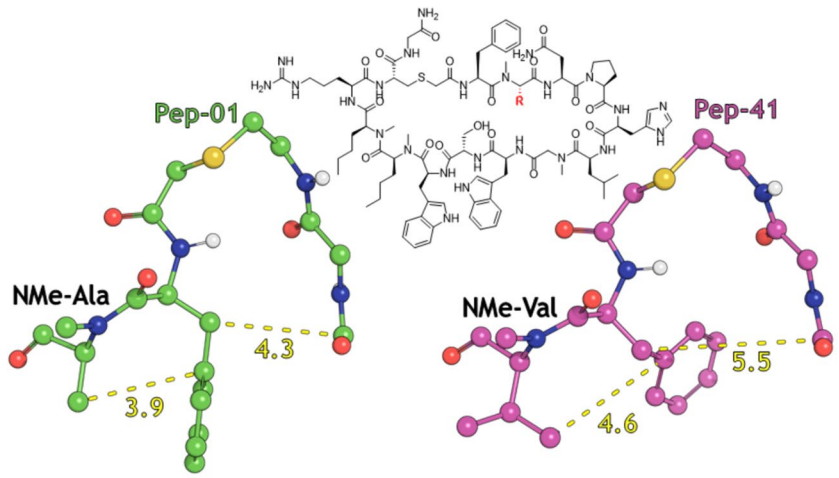


Fig. 3 Left: Top features for the PD-L1 full peptide SimFP data set. Green: Positive contribution, i.e., improving this interaction or maximizing this feature improves pIC_{50}. Red: Negative contribution, i.e., reducing this interaction or minimizing this feature improves pIC_{50}. Right: Lasso leave-one-molecule-out cross-validation (LOMO-CV) RMSE = 0.78 and $Q^2 = 0.36$. The parity plot shows ½ and 1 log error bands. Normalized strain energy is the top feature with a negative contribution. In other words, reducing strain helps with improving potency. Water-mediated interaction with Asn63 is identified to have the most detrimental contribution while water-mediated interaction with Val76 has the most positive contribution to the HTRF potency of these cyclic peptides to PD-L1



Peptide	Pep-01	Pep-41
R	CH_3	$CH(CH_3)_2$
HTRF pIC_{50}	8.1	6.0
Strain Energy	0.0721	0.0917
N63 W.B.	0.00	0.00
Y56 Hydro.	0.62	0.87
Y123 Pi-Pi	0.00	0.10
V76 W.B.	0.00	0.00
M115 Hydro.	0.37	0.28
Y123 Hydro.	0.49	0.46
V68 Hydro.	0.45	0.59
Y56 H-bond	0.00	0.00
Q66 W.B.	0.14	0.20

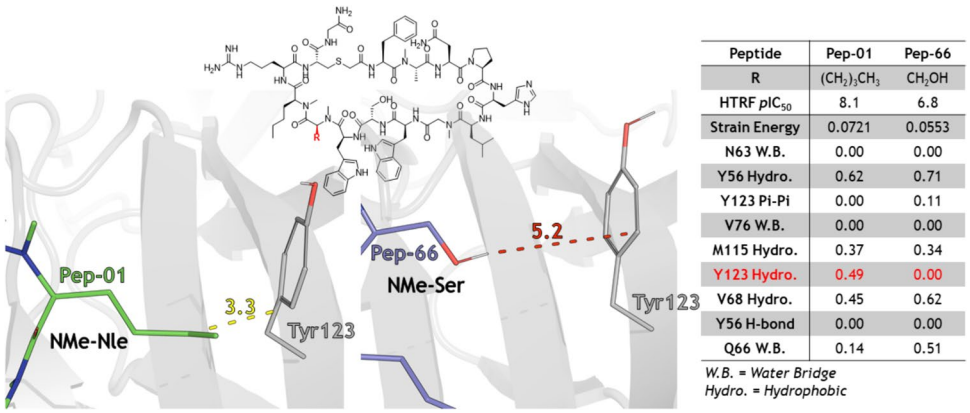
W.B. = Water Bridge
Hydro. = Hydrophobic

Fig. 4 Cluster representatives for matched molecular pairs Pep-01 and Pep-41 in the PD-L1 data set. The backbone conformation of Pep-41 is distorted compared to Pep-01, resulting in a much higher strain energy

for two additional heavy atoms. In a prospective peptide design exercise around this MMP, modifications would focus on reducing the strain energy of Pep-41 while retaining the SimFPs observed with MDFit.

MMPs with hydrophobic interactions differences

Pep-01 and Pep-66 differ only at position 11, where Pep-01 has NMe-Nle and Pep-66 has NMe-Ser. This mutation results in a significant loss in HTRF potency ($pIC_{50} = 8.1$ vs 6.8, respectively). Truncating the sidechain of Pep-01 results in a favorable reduction in strain energy but sacrifices a hydrophobic interaction with Tyr123 (Fig. 5). The attractive forces between Tyr123 and NMe-Nle fully liberate water in the binding interface, more fully optimizing the protein-peptide compatibility [67]. Smaller polar sidechains will not fully desolvate the binding site, compromising the binding affinity. This case exemplifies the importance of integrating SimFPs and metrics from rigid methods such as docking. Relying solely on strain energy for ligand optimization or prioritization would incorrectly rank Pep-66 higher than Pep-01. Without the high-throughput analysis of MD provided by MDFit, project teams could be misled, and optimization strategies may



Peptide	Pep-01	Pep-66
R	$(CH_2)_3CH_3$	CH_2OH
HTRF pIC_{50}	8.1	6.8
Strain Energy	0.0721	0.0553
N63 W.B.	0.00	0.00
Y56 Hydro.	0.62	0.71
Y123 Pi-Pi	0.00	0.11
V76 W.B.	0.00	0.00
M115 Hydro.	0.37	0.34
Y123 Hydro.	0.49	0.00
V68 Hydro.	0.45	0.62
Y56 H-bond	0.00	0.00
Q66 W.B.	0.14	0.51

W.B. = Water Bridge
Hydro. = Hydrophobic

Fig. 5 Cluster representatives for matched molecular pairs Pep-01 and Pep-66 in the PD-L1 data set. Pep-66 loses a hydrophobic attractive interaction with Tyr123 relative to Pep-01

lead to undesired outcomes. For peptide optimization in this MMP, designs would aim to recover the hydrophobic interaction in the Pep-01 MDFit SimFP while maintaining the lower strain energy observed for Pep-66.

Kullback–Leibler divergence for matched-pairs

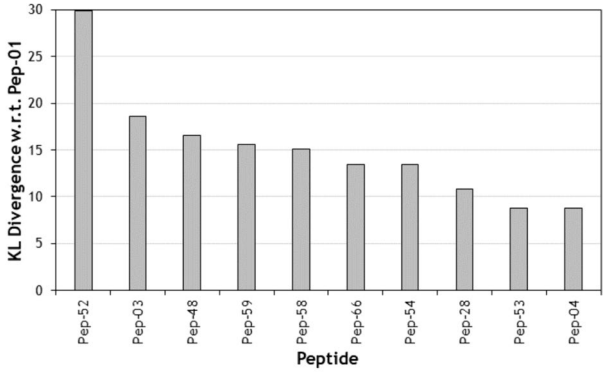
In some cases, differences in MD stabilities across the top features highlighted by the ML model do not fully explain the difference in potencies. Pep-52 features a beneficial water-mediated interaction with Val76 which is not observed for Pep-01 (importance = +0.64) as well as an amplified detrimental water-mediated interaction with Gln66 (importance = –0.38). All other SimFP features were remarkably similar between the two peptides. Based on only these features, one would expect Pep-52 to have equal or slightly better HTRF potency compared to Pep-01. However, Pep-52 was about fivefold less potent than Pep-01.

The Kullback–Leibler divergence (KL divergence, relative entropy [68]) between SimFPs offers an alternate quantification strategy that characterizes differences across all the features in the SimFPs into a single dimensional quantity. SimFP of Pep-01 is treated as the reference and KL divergence for all the other peptides in the series was calculated relative to Pep-01. KL divergence identified Pep-52 to have the most divergent SimFP compared to Pep-01 (29.9; Fig. 6A) prompting further investigation.

The difference between the raw SimFPs ($|SimFP_{Pep\text{-}52} – SimFP_{Pep\text{-}01}|$) identified the water-mediated interaction with Gln66 as the single most divergent SimFP feature across all three repetitions of Pep-01 and Pep-52. The detrimental water-mediated interaction between Pep-52 and Gln66 for the individual repetition SimFPs were 80%, 61%, and 55% (Fig. 6B; trajectory 3, 2, 1, respectively). In contrast, Pep-01 featured this interaction a mere 42% in trajectory 2 and never registered (0%) in trajectories 1 and 3.

Visualizing the representative clusters for Pep-01 and Pep-52 revealed the backbone carbonyl of Pro4 in Pep-52 forms a water-mediated hydrogen bond with the backbone

A

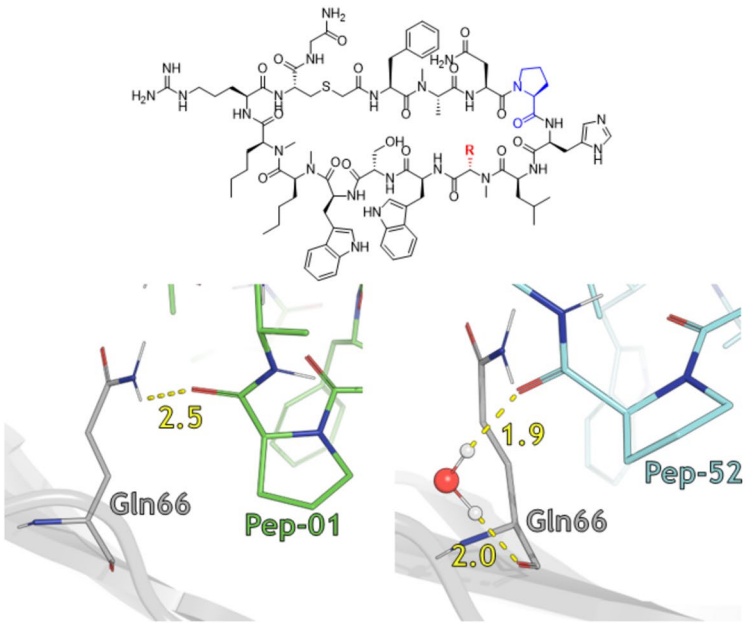


B

Peptide	Q66 W.B. SimFP Traj. 1	Traj. 2	Traj. 3
Pep-01	0.00	0.42	0.00
Pep-52	0.55	0.61	0.80

Fig. 6 **A** The top 10 most divergent SimFPs by KL divergence relative to Pep-01. **B** Differences in water-mediated interactions with Gln66 across three runs of MD. Pep-01 features less water-mediated interaction with Gln66 compared to Pep-52 indicating a tighter binding, compared to Pep-52 where water has seeped into the pocket

carbonyl of Gln66 (Fig. 7). For Pep-01, the same backbone carbonyl of Pro4 hydrogen bonds directly with the sidechain of Gln66. Water infiltration characterizes protein-peptide incompatibility for Pep-52, explaining the drop in HTRF potency relative to Pep-01 (pIC_{50} = 7.6 vs 8.1, respectively). While incompatibility may be observed tangentially in computational methods that treat proteins as rigid bodies, direct observation of water infiltration at a specific residue from dynamic models focuses the project team on an area of the ligand for further optimization. In this case, a deep dive into ML feature importance, KL divergence, and raw SimFPs helped differentiate the peptide's behavior in the binding pocket and explain the difference in potency.



Peptide	Pep-01	Pep-52
R	H	$(CH_2)_2NH_2$
HTRF pIC_{50}	8.1	7.6
Strain Energy	0.0721	0.0729
N63 W.B.	0.00	0.11
Y56 Hydro.	0.62	0.77
Y123 Pi-Pi	0.00	0.00
V76 W.B.	0.00	0.50
M115 Hydro.	0.37	0.47
Y123 Hydro.	0.49	0.44
V68 Hydro.	0.45	0.29
Y56 H-bond	0.00	0.00
Q66 W.B.	0.14	0.65

W.B. = Water Bridge
Hydro. = Hydrophobic

Fig. 7 Cluster representatives for matched molecular pairs Pep-01 and Pep-52 in the PD-L1 data set. Pro4 engages Gln66 through a direct hydrogen bond (Pep-01) or a water-mediated hydrogen bond (Pep-52). The water infiltration in the Pep-52 simulations provides a possible explanation for the difference in potency relative to Pep-01

Small molecule CDK9 inhibitors

Although there is some semblance of correlation ($R^2 = 0.1$) between docked pose G lide scores and enzyme pIC_{50} (SI Figure S6), ML learning with SimFP was useful in delineating interesting SAR trends, particularly those involving water-mediated interactions that were otherwise missed in docking. The top ten features (weights with the largest absolute value) of the CDK9 data set are reported in Fig. 8, left. A hydrophobic interaction with Phe103 (Fig. 1C, Fig. 8) was the most detrimental (negative weight) and a hydrophobic interaction with Ile25 was the most beneficial (positive weight) SimFP to potency. Select Match Molecular Pair (MMP) cases will be described herein using the feature selection to explain SAR.

MMPs with hydrophobic interactions differences

Compound 24 and Compound 22 differ only around the pyridinone core, where Compound 24 is an isopropylpyridine ring and compound 22 has a methoxy-methylpyridinone. This core modification results in a significant loss in potency ($pIC_{50} = 8.5$ vs 4.5, respectively). Unsubstituted pyridinone of Compound 24 results in a favorable reduction in hydrophobic contact with Phe103 in favor of a beneficial pi-pi stacking interaction (Fig. 9). For small molecule optimization in this MMP, designs would aim to remove the hydrophobic interaction in Compound 22.

MMPs with water infiltration

Compound 24 and Compound 30 differ only around the pyridinone core R-group, where Compound 24 is an isopropyl and Compound 30 has a tetrahydropyran. This solvent-exposed modification results in a significant loss in potency ($pIC_{50} = 8.5$ vs 5.4, respectively). The smaller R-group of Compound 24 results in better protein–ligand compatibility than the bulkier R-group of Compound 30. The pyridinone core of Compound 30 distorts and drifts in the binding site, allowing water infiltration, observed as a disfavored water-mediated interaction with Asp167 (Fig. 10). For small molecule optimization in this MMP, designs would aim to probe the R-group size tolerability and the effects on protein–ligand compatibility.

Simulation length

To enable efficient rank-ordering of peptide designs using SimFPs prospectively, it is important to also assess simulation convergence. Root Mean Square Deviation (RMSD) of the ligand conformations relative to the protein pocket is an often discussed metric to estimate convergence. However, as shown in Fig. 11, RMSD plots are not always useful in estimating how long a simulation needs to be for full convergence. Instead, the divergence of SimFPs from different time intervals relative to the full simulation trajectory (100 ns) can be used to estimate simulation convergence (Fig. 11). For the reference Pep-01 in the PD-L1 dataset

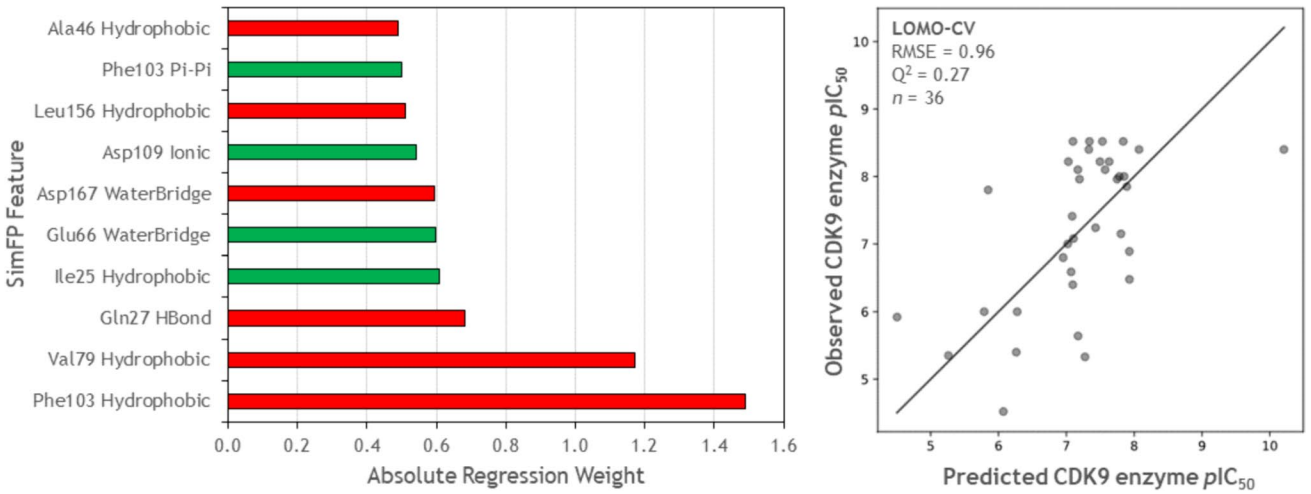
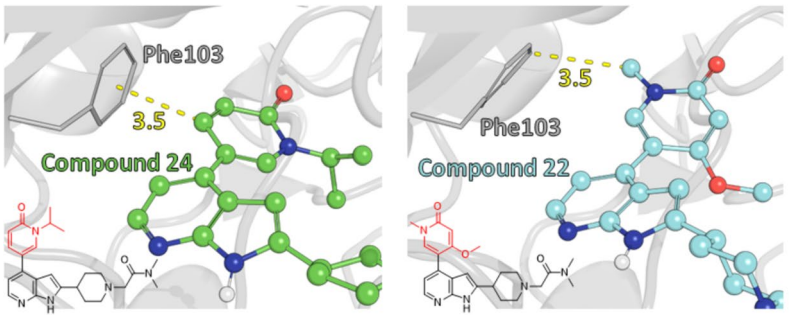
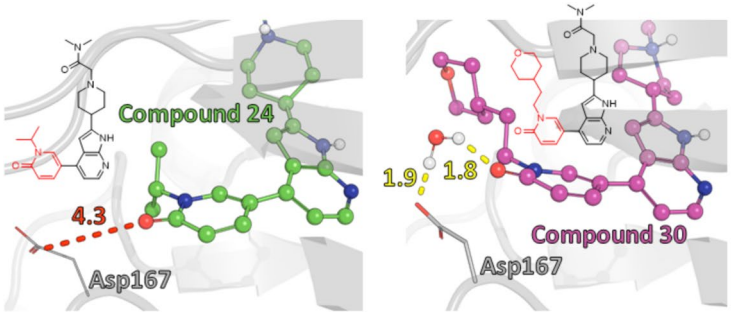


Fig. 8 Left: Top features for the full CDK9 SimFP data set. Green: Positive contribution, i.e., improving this interaction or maximizing this feature improves pIC_{50}. Red: Negative contribution, i.e., reducing this interaction or minimizing this feature improves pIC_{50}. Right: Lasso leave-one-molecule-out cross-validation (LOMO-CV) RMSE = 0.96 and $Q^2 = 0.27$. Hydrophobic contact with Phe103 is the top feature with a negative contribution. In other words, reducing the interaction prevalence helps with improving potency. Hydrophobic interaction with Ile25 is identified to have the most positive contribution to the HTRF potency of the CDK9 inhibitors



Compound	24	22
CDK9 pIC_{50}	8.5	4.5
Phe103 Hydro.	0.13	0.65
Val79 Hydro.	0.06	0.10
Gln27 H-bond	0.00	0.10
Ile25 Hydro.	0.10	0.22
Glu66 W.B.	0.00	0.00
Phe103 Pi-Pi	0.85	0.17

Fig. 9 Cluster representatives for matched molecular pairs Compound 24 and Compound 22 in the CDK9 data set. Pyridinone engages Phe103 through a direct pi-pi stacking interaction (Compound 24) or hydrophobic contact (Compound 22). The replacement of the pi–pi interaction with a hydrophobic interaction during the simulations provides a possible explanation for the difference in potency relative to Compound 24



Compound	24	30
CDK9 pIC_{50}	8.5	5.4
Phe103 Hydro.	0.13	0.10
Val79 Hydro.	0.06	0.05
Gln27 H-bond	0.00	0.19
Ile25 Hydro.	0.10	0.18
Glu66 W.B.	0.00	0.10
Asp167 W.B.	0.01	0.79

Fig. 10 Cluster representatives for matched molecular pairs Compound 24 and Compound 30 in the CDK9 data set. Piperidinone engages Phe103 through a water-mediated interaction (Compound 30) or does not engage Asp167 (Compound 24). The water infiltration in the binding site during the simulations provides a possible explanation for the difference in potency relative to Compound 24

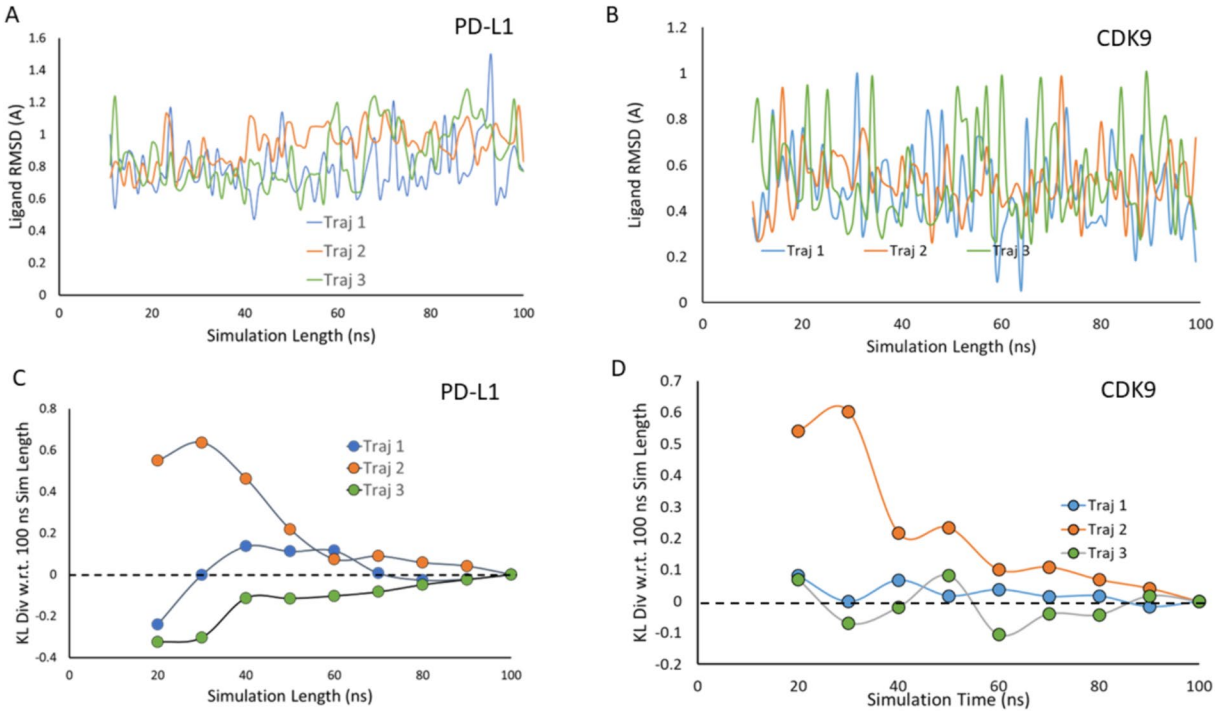


Fig. 11 **A**, **B**: Heavy-atom RMSD of Pep-01 in PD-L1 and Compound 38 in CDK9 data set respectively relative to the protein pocket throughout the three 100 ns MD repetitions. **C**, **D**: KL divergence of Pep-01 & Compound 38 SimFPs relative to the full trajectory (100 ns) shows that simulations converge at 70 ns

and Compound 38 in the CDK9 dataset, SimFPs converge at 70 ns for all three MD trajectory repetitions. Therefore, 100 ns MD trajectories can be assumed to fully characterize relevant protein–ligand dynamics, and ML models can rank-order designs using the SimFPs.

Conclusions

We have presented a new high-throughput workflow for setting up, running, and analyzing molecular dynamics simulations for a library of ligands. MDFit produces compiled simulation fingerprints (SimFPs) for users to decipher critical protein–ligand interactions and rank-order ligands based on compatibility. Application of the MDFit workflow to a data set of 61 peptides bound to PD-L1 & 39 small-molecule inhibitors bound to CDK9 resulted in the discovery of several SimFPs critical for HTRF potency & binding. Matched molecular pairs were explored to highlight the utility of SimFPs when combined with ML techniques. KL divergence offers an attractive alternative to explain potency differences otherwise not evident in the top ML features.

The stability of pocket interactions from MD simulations characterizes the enthalpy of binding into the protein pocket. Conformational entropy is included via pre-calculated strain of the docked pose [53] in the SimFP. Through sufficient sampling of each ligand in the binding pocket, ML models trained on these SimFPs account for all important thermodynamic events and therefore have reasonable accuracy of predictions of binding affinity. Unlike relative free energy perturbation [69] approaches that have limitations based on ligand size [54] and chemical similarity [70], SimFP-based ML models for potency assessment are less likely to have either of these constraints. Future version releases will support other MD engines (OpenMM [71], GROMACS [72]) and force-fields (OpenFF [73]), add more information into SimFPs [73], and additional analysis via machine learning approaches. While the current version uses Schrodinger's native simulation interaction analysis APIs for Desmond trajectories, for SimFPs with OpenMM/GROMACS trajectories we will integrate ProLif [48] into our workflow. The MDFit workflow is expected to be useful for characterizing pocket dynamics of multiple modalities, including small molecules, peptides, PROTACs, and molecular glues to drive drug discovery projects moving forward.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10822-024-00564-2.

Acknowledgements The authors wish to thank Scott Hollingsworth, Ahmet Mentes, Alexios Koutsoukas, and Brian Claus at Bristol Myers Squibb for helpful discussions around the development and implementation of MDFit and machine-learning models using SimFPs. The authors thank Bristol Myers Squibb for sponsoring a summer internship for A.S. (Rutgers University, New Brunswick, New Jersey). The authors also wish to thank Ajay Jain and Ann Cleves at Optibrium, Ltd. for discussions around the inclusion of strain energy for the PD-L1 SimFP data set.

Author contributions All authors participated in the research and in the preparation of and final review of the manuscript.

Funding The authors have no outside funding to declare.

Data availability The workflow is freely available for download from GitHub repository: https://github.com/brueckna2020/MDFit. This repository contains the PD-L1 dataset including protein model, ligand docked poses, MD trajectories as well as the ML model discussed in the manuscript in the examples directory.

Declarations

Competing interests The authors declare no competing interests.

Ethical approval Not applicable.

References

1. Bajad NG, Rayala S, Gutti G, Sharma A, Singh M, Kumar A, Singh SK (2021) Systematic review on role of structure based drug design (SBDD) in the identification of anti-viral leads against SARS-Cov-2. Curr Res Pharmacol Drug Discov 2:100026
2. Jorgensen WL (2004) The many roles of computation in drug discovery. Science 303(5665):1813–1818
3. De Vivo M (2011) Bridging quantum mechanics and structure-based drug design. Front Biosci 16(5):1619–1633
4. Schneider G (2010) Virtual screening: an endless staircase? Nat Rev Drug Discov 9(4):273–276
5. Ain QU, Aleksandrova A, Roessler FD, Ballester PJ (2015) Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdiscip Rev Comput Mol Sci 5(6):405–424
6. Jorgensen WL (2009) Efficient drug lead discovery and optimization. Acc Chem Res 42(6):724–733
7. Jakhar R, Dangi M, Khichi A, Chhillar AK (2020) Relevance of molecular docking studies in drug designing. Curr Bioinform 15(4):270–278
8. De Vivo M, Masetti M, Bottegoni G, Cavalli A (2016) Role of molecular dynamics and related methods in drug discovery. J Med Chem 59(9):4035–4061
9. Hollingsworth SA, Dror RO (2018) Molecular dynamics simulation for all. Neuron 99(6):1129–1143
10. Beckstein O, Sansom MS (2006) A hydrophobic gate in an ion channel: the closed state of the nicotinic acetylcholine receptor. Phys Biol 3(2):147
11. Marino KA, Shang Y, Filizola M (2018) Insights into the function of opioid receptors from molecular dynamics simulations of available crystal structures. Br J Pharmacol 175(14):2834–2845
12. Freddolino PL, Arkhipov AS, Larson SB, McPherson A, Schulten K (2006) Molecular dynamics simulations of the complete satellite tobacco mosaic virus. Structure 14(3):437–449
13. Li J, Flick F, Verheugd P, Carloni P, Lüscher B, Rossetti G (2015) Insight into the mechanism of intramolecular inhibition of the catalytic activity of sirtuin 2 (SIRT2). PLoS One 10(9):e0139095
14. Pantsar T, Rissanen S, Dauch D, Laitinen T, Vattulainen I, Poso A (2018) Assessment of mutation probabilities of KRAS G12 missense mutants and their long-timescale dynamics by atomistic molecular simulations and Markov state modeling. PLoS Comput Biol 14(9):e1006458
15. Salo-Ahen OM, Alanko I, Bhadane R, Bonvin AM, Honorato RV, Hossain S, Juffer AH, Kabedev A, Lahtela-Kakkonen M, Larsen AS (2020) Molecular dynamics simulations in drug discovery and pharmaceutical development. Processes 9(1):71
16. Borhani DW, Shaw DE (2012) The future of molecular dynamics simulations in drug discovery. J Comput Aided Mol Des 26:15–26
17. Coop A, MacKerell A (2002) The future of opioid analgesics. Am J Pharm Educ 66(2):153–156
18. Healy JR, Bezawada P, Shim J, Jones JW, Kane MA, MacKerell AD Jr, Coop A, Matsumoto RR (2013) Synthesis, modeling, and pharmacological evaluation of UMB 425, a mixed μ agonist/δ antagonist opioid analgesic with reduced tolerance liabilities. ACS Chem Neurosci 4(9):1256–1266
19. Shaw DE, Grossman J, Bank JA, Batson B, Butts JA, Chao JC, Deneroff MM, Dror RO, Even A, Fenton CH (2014) Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In: SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, New York, pp 41–53
20 Harvey MJ, De Fabritiis G (2015) AceCloud: molecular dynamics simulations in the cloud. ACS Publications, Washington
21. Kondratyuk N, Nikolskiy V, Pavlov D, Stegailov V (2021) GPU-accelerated molecular dynamics: State-of-art software performance and porting from Nvidia CUDA to AMD HIP. Int J High Perform Comput Appl 35(4):312–324
22. Turalija M, Petrović M, Kovačić B (2022) Towards general-purpose long-timescale molecular dynamics simulation on exascale supercomputers with data processing units. In: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), IEEE, New York, pp 300–306
23. Harvey MJ, Giupponi G, Fabritiis GD (2009) ACEMD: accelerating biomolecular dynamics in the microsecond time scale. J Chem Theory Comput 5(6):1632–1639
24. Alam S, Varettoa U (2014) GROMACS on Hybrid CPU-GPU and CPU-MIC clusters: preliminary porting experiences, results and next steps
25. Bergdorf M, Robinson-Mosher A, Guo X, Law KH, Shaw DE (2021) Desmond/GPU performance as of April 2021. DE Shaw Research, Tech. Rep. DESRES/TR–2021-01
26. Mosalaganti S, Obarska-Kosinska A, Siggel M, Taniguchi R, Turoňová B, Zimmerli CE, Buczak K, Schmidt FH, Margiotta E, Mackmull M-T (2022) AI-based structure prediction empowers integrative structural analysis of human nuclear pores. Science 376(6598):eabm9506
27. Nawrocki G, Im W, Sugita Y, Feig M (2019) Clustering and dynamics of crowded proteins near membranes and their influence on membrane bending. Proc Natl Acad Sci 116(49):24562–24567
28. Pezeshkian W, König M, Wassenaar TA, Marrink SJ (2020) Backmapping triangulated surfaces to coarse-grained membrane models. Nat Commun 11(1):2296
29. Stevens JA, Grünewald F, van Tilburg PM, König M, Gilbert BR, Brier TA, Thornburg ZR, Luthey-Schulten Z, Marrink SJ (2023) Molecular dynamics simulation of an entire cell. Front Chem 11:1106495
30. Luthey-Schulten Z, Thornburg ZR, Gilbert BR (2022) Integrating cellular and molecular structures and dynamics into whole-cell models. Curr Opin Struct Biol 75:102392

31. Perilla JR, Goh BC, Cassidy CK, Liu B, Bernardi RC, Rudack T, Yu H, Wu Z, Schulten K (2015) Molecular dynamics simulations of large macromolecular complexes. Curr Opin Struct Biol 31:64–74
32. Skånberg R, Linares M, König C, Norman P, Jönsson D, Hotz I, Ynnerman A (2018) VIA-MD: Visual interactive analysis of molecular dynamics. In: MolVa@ EuroVis, pp 19–27
33. Magarkar A (2023) MD-Simba. https://github.com/aniketsh/MD-SIMBA-Public (accessed) 2020
34 Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, Papadopoulos K, Patronov A (2020) REINVENT 2.0: an AI tool for de novo drug design. J Chem Inf Model 60(12):5918–5922
35. Glaser J, Vermaas JV, Rogers DM, Larkin J, LeGrand S, Boehm S, Baker MB, Scheinberg A, Tillack AF, Thavappiragasam M (2021) High-throughput virtual laboratory for drug discovery using massive datasets. Int J High Perform Comput Appl 35(5):452–468
36. Scientific O (2019) GigaDocking™—structure based virtual screening of over 1 billion molecules webinar
37. Rogers DM, Agarwal R, Vermaas JV, Smith MD, Rajeshwar RT, Cooper C, Sedova A, Boehm S, Baker M, Glaser J (2023) SARS-CoV2 billion-compound docking. Sci Data 10(1):173
38. Boby ML, Fearon D, Ferla M, Filep M, Koekemoer L, Robinson MC, Consortium CM, Chodera JD, Lee AA, London N (2023) Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. Science 382(6671):eabo7201
39. CourniaAllenSherman ZBW (2017) Relative binding free energy calculations in drug discovery: recent advances and practical considerations. J Chem Inf Model 57(12):2911–2937
40. Merz KM Jr (2010) Limits of free energy computation for protein– ligand interactions. J Chem Theory Comput 6(5):1769–1776
41. Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA, Klepeis JL, Kolossvary I, Moraes MA, Sacerdoti FD (2006) Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, pp 84-es
42. Lu C, Wu C, Ghoreishi D, Chen W, Wang L, Damm W, Ross GA, Dahlgren MK, Russell E, Von Bargen CD (2021) OPLS4: Improving force field accuracy on challenging regimes of chemical space. J Chem Theory Comput 17(7):4291–4300
43. Gebhardt J, Kiesel M, Riniker S, Hansen N (2020) Combining molecular dynamics and machine learning to predict self-solvation free energies and limiting activity coefficients. J Chem Inf Model 60(11):5319–5330
44. Riniker S (2017) Molecular dynamics fingerprints (MDFP): machine learning from MD data to predict free-energy differences. J Chem Inf Model 57(4):726–741
45. Baumgartner MP, Zhang H (2020) Building admiral, an automated molecular dynamics and analysis platform. ACS Med Chem Lett 11(11):2331–2335
46. Liu K, Watanabe E, Kokubo H (2017) Exploring the stability of ligand binding modes to proteins by molecular dynamics simulations. J Comput Aided Mol Des 31:201–211
47. Bouysset C, Fiorucci S (2021) ProLIF: a library to encode molecular interactions as fingerprints. J Cheminform 13(1):72
48. Kythreotou A, Siddique A, Mauri FA, Bower M, Pinato DJ (2018) PD-L1. J Clin Pathol 71(3):189–194
49. Yi M, Zheng X, Niu M, Zhu S, Ge H, Wu K (2022) Combination strategies with PD-1/PD-L1 blockade: current advances and future directions. Mol Cancer 21(1):28
50. Barlaam B, De Savi C, Dishington A, Drew L, Ferguson AD, Ferguson D, Gu C, Hande S, Hassall L, Hawkins J (2021) Discovery of a series of 7-azaindoles as potent and highly selective CDK9 inhibitors for transient target engagement. J Med Chem 64(20):15189–15213
51. Borowczak J, Szczerbowski K, Ahmadi N, Szylberg Ł (2022) CDK9 inhibitors in multiple myeloma: a review of progress and perspectives. Med Oncol 39(4):39
52. Jain AN, Brueckner AC, Jorge C, Cleves AE, Khandelwal P, Cortes JC, Mueller L (2023) Complex peptide macrocycle optimization: combining NMR restraints with conformational analysis to guide structure-based and ligand-based design. J Comput Aided Mol Des 37:519–535
53. Wallraven K, Holmelin FL, Glas A, Hennig S, Frolov AI, Grossmann TN (2020) Adapting free energy perturbation simulations for large macrocyclic ligands: how to dissect contributions from direct binding and free ligand flexibility. Chem Sci 11(8):2269–2276
54. Niu B, Appleby TC, Wang R, Morar M, Voight J, Villaseñor AG, Clancy S, Wise S, Belzile J-P, Papalia G (2019) Protein footprinting and X-ray crystallography reveal the interaction of PD-L1 and a macrocyclic peptide. Biochemistry 59(4):541–551
55. Rostkowski M, Olsson MH, Søndergaard CR, Jensen JH (2011) Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. BMC Struct Biol 11:1–6
56. Miller MM, Mapelli C, Allen MP, Bowsher MS, Boy KM, Gillis EP, Langley DR, Mull E, Poirier MA, Sanghvi N (2014) Macrocyclic inhibitors of the PD-1/PD-L1 and CD80 (B7-1)/PD-L1 protein/protein interactions. Google Patents
57. Jiao L, Dong Q, Zhai W, Zhao W, Shi P, Wu Y, Zhou X, Gao Y (2022) A PD-L1 and VEGFR2 dual targeted peptide and its combination with irradiation for cancer immunotherapy. Pharmacol Res 182:106343
58. Repasky MP, Murphy RB, Banks JL, Greenwood JR, Tubert-Brohman I, Bhat S, Friesner RA (2012) Docking performance of the glide program as evaluated on the Astex and DUD datasets: a complete set of glide SP results and selected results for a new scoring function integrating WaterMap and glide. J Comput Aided Mol Des 26(6):787–799
59. Watts KS, Dalal P, Tebben AJ, Cheney DL, Shelley JC (2014) Macrocycle conformational sampling with MacroModel. J Chem Inf Model 54(10):2680–2696
60. Wu Y, Tepper HL, Voth GA (2006) Flexible simple point-charge water model with improved liquid-state properties. J Chem Phys. https://doi.org/10.1063/1.2136877
61. Berendsen HJ, Hayward S (2000) Collective protein dynamics in relation to function. Curr Opin Struct Biol 10(2):165–169
62. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4(2):187–217
63 Lindahl E, Hess B, Van Der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7(8):306–317
64. Jorgensen WL, Schyman P (2012) Treatment of halogen bonding in the OPLS-AA force field: application to potent anti-HIV agents. J Chem Theory Comput 8(10):3895–3901
65. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
66 Klebe G (2015) Protein-ligand interactions as the basis for drug action. Multifaceted roles of crystallography in modern drug discovery. Springer, New York, pp 83–92
67. Kullback S (1951) Kullback-leibler divergence
68. Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J (2015) Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. J Am Chem Soc 137(7):2695–2703

69. Schindler CE, Baumann H, Blum A, Böse D, Buchstaller H-P, Burgdorf L, Cappel D, Chekler E, Czodrowski P, Dorsch D (2020) Large-scale assessment of binding free energy calculations in active drug discovery projects. J Chem Inf Model 60(11):5457–5474
70. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang L-P, Simmonett AC, Harrigan MP, Stern CD (2017) OpenMM 7: rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol 13(7):e1005659
71. Hess B, Kutzner C, Van Der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4(3):435–447
72 Lim VT, Hahn DF, Tresadern G, Bayly CI, Mobley DL (2020) Benchmark assessment of molecular geometries and energies from small molecule force fields. F1000Res 9:Chem Inf Sci-1390
73. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. J Comput Chem 32(10):2319–2327