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]

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

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

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 using MD. We present an automated workflow that streamlines setting up, running, and analyzing Desmond MD simulations. The workflow takes a library of pre-docked ligands and a 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. SimFP from a ligand library can also be used to build machine learning (ML) models that can 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 a case study on how SimFP helps delineate structure-activity relationship (SAR) trends and explain potency differences across matched-molecular pairs of cyclic peptides targeting the PD-L1 protein.

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 relation between pocket dynamics and function of several different classes of proteins including transmembrane receptors like ion channels[10], opioid receptors[11], viral capsids[12], sirtuins[13], and RAS[14] family proteins among others. These MD 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 a typical optimization campaign. Moreover, several recent works have explored 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 generative ML or at the end of large-library docking exercises is likely to improve the accuracy of predictions and enrichment of hits from these workflows.

We present an automated workflow (MDFit) that streamlines setting up, running, and analyzing Desmond[25, 42] MD simulations 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 the MD trajectories for 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 (SimFP) 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.

We show an application of MDFit for assessing a set of cyclic peptides that target PD-L1 with therapeutical potential as anticancer agents[44, 45]. PD-L1 binds to PD-1 at an elongated ß-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 ß-sheet interface between PD-1 and PD-L1 (Figure 1, left). 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 (Figure 1, right).

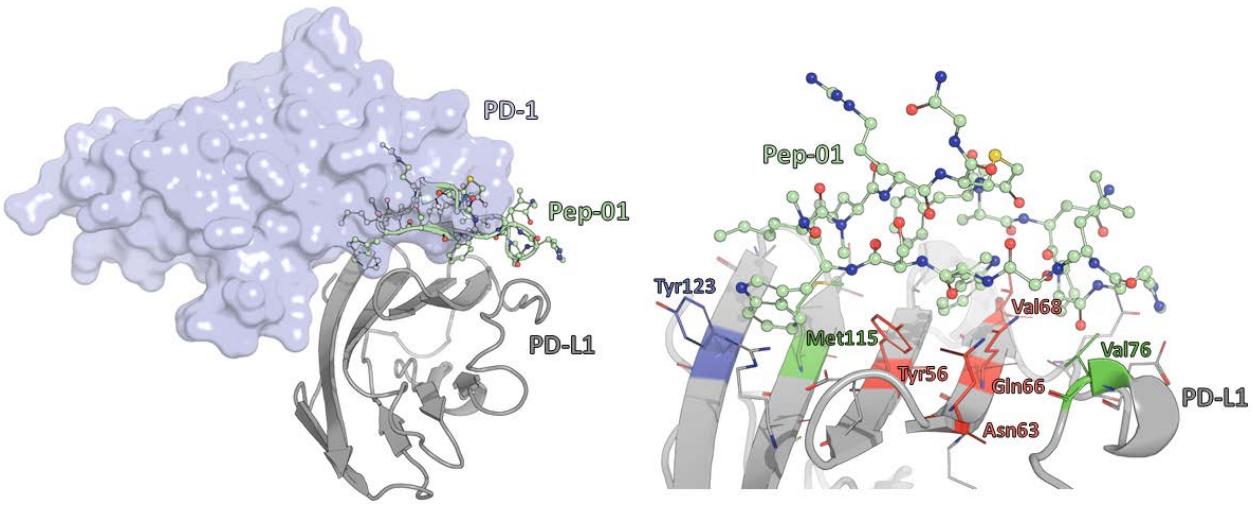


Figure 1. (Left) Overlay of 4ZQK and 6PV9 crystal structures showing Pep-01 binds to the ß-sheet interface of PD-L1 to block PD-1 binding. (Right) Peptide binding interface with PD-L1. All residues within 5 Å of Pep-01 are shown with critical residues determined by ML models (*vide infra*) shown in red (detrimental), green (beneficial), or blue (detrimental or beneficial).

Previous studies have shown a strong correlation between strain measured in these peptides to mimic PD-1 binding and their potency through docking of extensively sampled conformations of the peptides[46]. The extremely large number of rotatable dihedrals with these cyclic peptides makes relative free-energy perturbation methods for assessing potency and pocket dynamics untenable[47]. 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.

Methods

The MDFit workflow (Figure 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 discussed below, protein from PDB 6PV9[48] was used as the starting protein conformation. Protonation states of protein residues were determined using PropKa[49] and 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[50] and sixty of its analogs as described here[50] 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[46, 50, 51]. Top poses from the docked conformer ensembles from the previous study[46] were used as starting conformations for MDFit described here.

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[52] 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 of either 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[53] is used to model water. All simulations are run using the Desmond[25] engine with Schrodinger suite version 2022-2.
    b. Solvated protein-ligand systems are relaxed before the production MD simulations. Initially, the entire system is equilibrated for 100ps using the NVT brownian dynamics at T=10K, 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 24ps using a Berendsen[54] thermostat with pressure gradually dropping from 50 to 2 bars through an NPT dynamics run.

5) Production: This is followed by production MD simulations using NPT dynamics and positional restraints removed. 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. For the PD-L1 dataset, each peptide-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[55-57]. 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[46] 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 the PD-L1 dataset, the first 10ns (100 frames) 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[58] (see Supporting Information; Table S2, Figure 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.

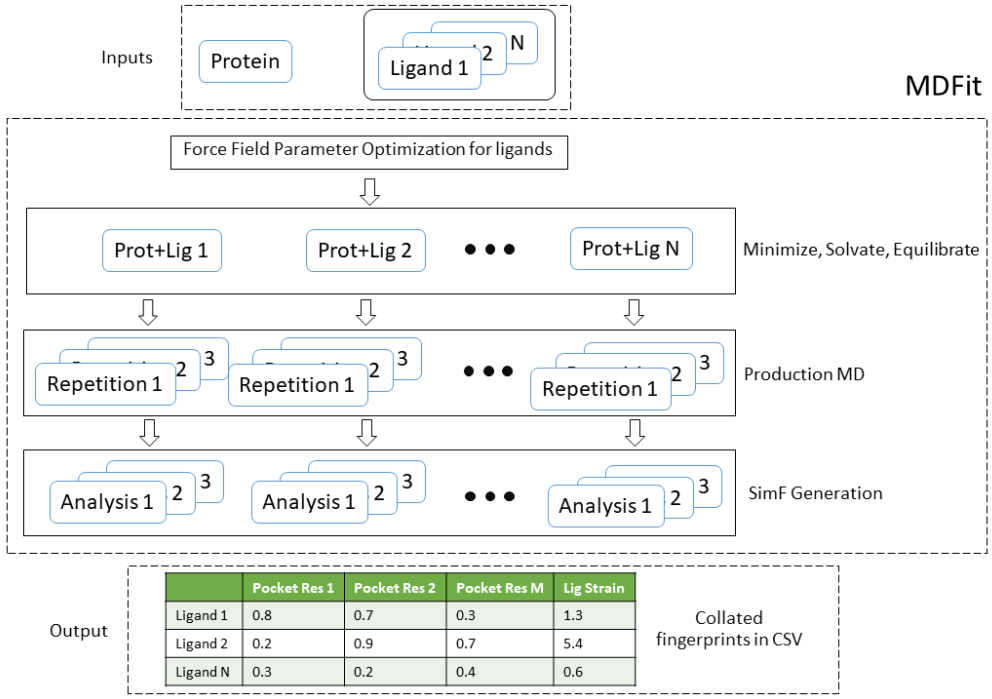


Figure 2. MDFit workflow takes a library of ligands with reasonable starting poses in a protein pocket, runs MD, and generates collated SimFP for easy analysis of the stability of all ligands in the protein pocket across MD trajectories.

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 Selection

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[46] as features. While the model performance was modest (Figure 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 Figure 3, left. We note that along with interaction stability fingerprints that come from MDFit, pre-computed strain energies[46] were included as an additional feature of SimFP. Strain energy remains the standout feature, consistent with previous studies[46], 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.

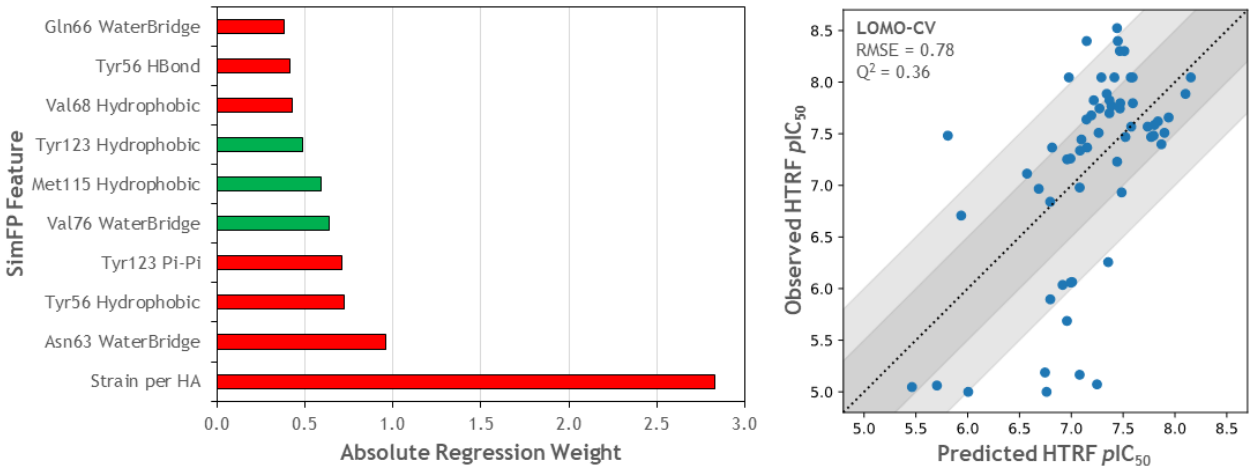
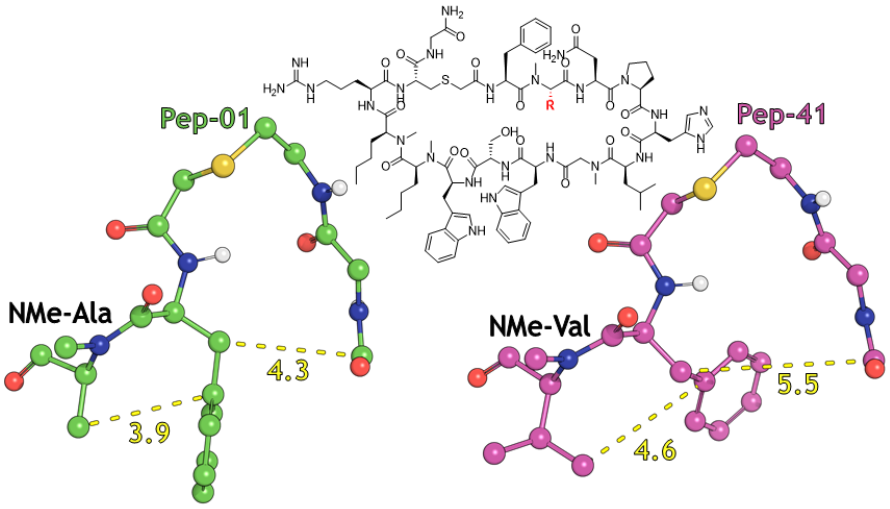


Figure 3. Left : Top features for the full peptide SimFP data set are shown in left plot. Green : Positive contribution, i.e., improving this interaction or maximizing this feature improves pIC_{50}. Red : Negative contribution, i.e., reducing this interaction or minimizing this feature improves pIC_{50}. 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 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.

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 (Figure 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 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.



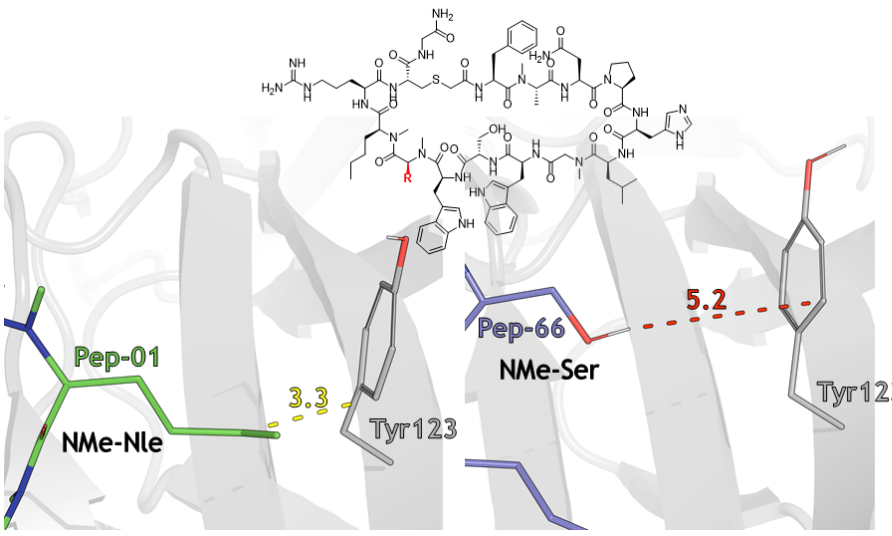
Peptide	Pep-01	Pep-41
R	CH_3	$CH(CH_3)_2$
HTRF pIC_{50}	8.1	6.0
Strain Energy	0.0721	0.0917
N63 W.B.	0.00	0.00
Y56 Hydro.	0.62	0.87
Y123 Pi-Pi	0.00	0.10
V76 W.B.	0.00	0.00
M115 Hydro.	0.37	0.28
Y123 Hydro.	0.49	0.46
V68 Hydro.	0.45	0.59
Y56 H-bond	0.00	0.00
Q66 W.B.	0.14	0.20

W.B. = Water Bridge
Hydro. = Hydrophobic

Figure 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.

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 (Figure 5). The attractive forces between Tyr123 and NMe-Nle fully liberate water in the binding interface, more fully optimizing the protein-peptide compatibility[59]. 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 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.



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

Figure 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.

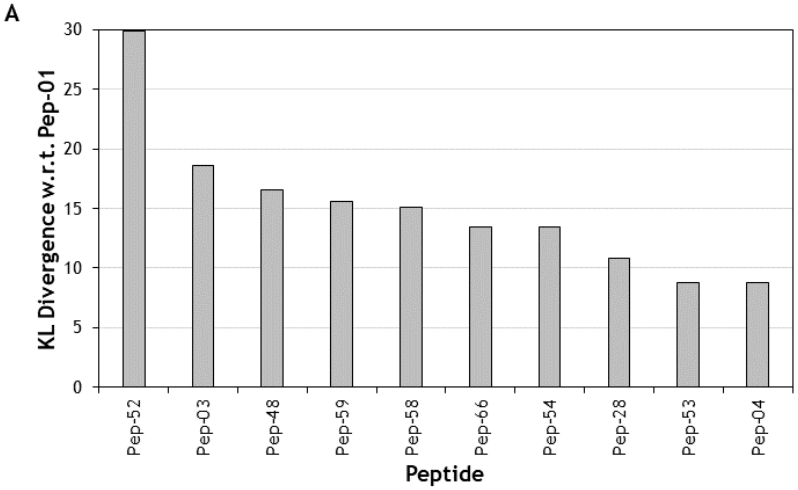
Kullback-Leibler Divergence for Matched-pairs

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 5 fold less potent than Pep-01.

The Kullback-Leibler divergence (KL divergence, relative entropy[60]) 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; Figure 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% (Figure 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.

A



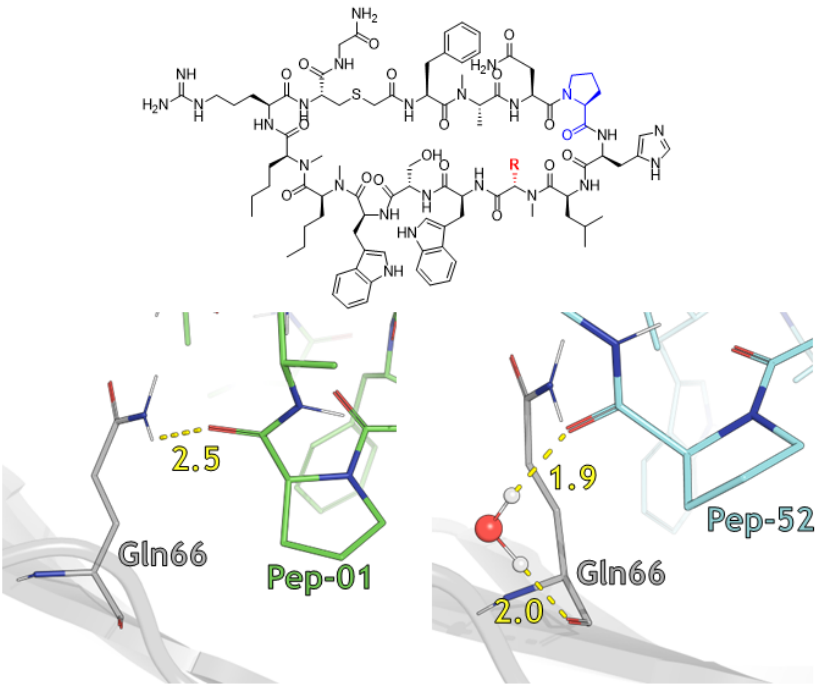
B

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

Figure 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.

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 carbonyl of Gln66 (Figure 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

Figure 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.

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 Figure 8A,

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 (Figure 8B). For the reference Pep-01, 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.

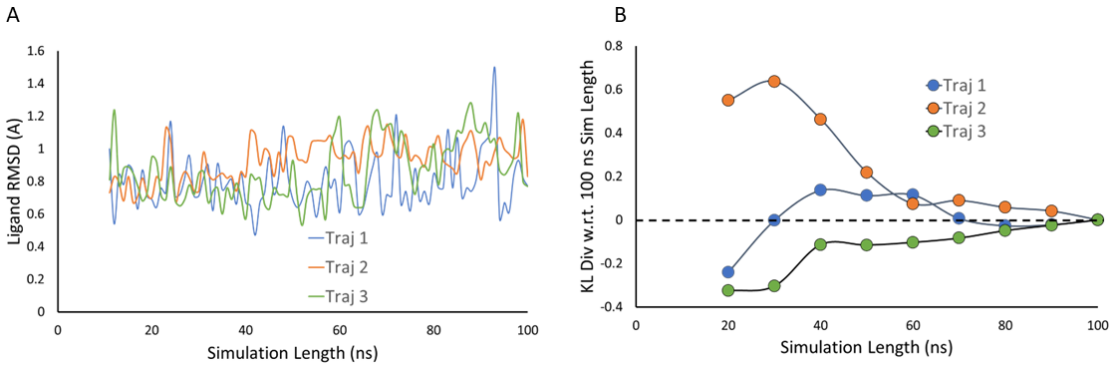


Figure 8. (A) Heavy-atom RMSD of Pep-01 relative to the protein pocket throughout the three 100 ns MD repetitions. (B) KL divergence of SimFPs relative to the full trajectory (100 ns) shows that simulations converge at 70 ns.

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 resulted in the discovery of

several SimFPs critical for HTRF potency. 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 that are 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[46] 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[61] approaches that have limitations based on ligand size[47] and chemical similarity[62], 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[63], GROMACS[64]) and force-fields (OpenFF[65]), add more information into SimFPs[66], and additional analysis via machine learning approaches. 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.

Acknowledgments

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 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.

References

(1) Bajad, N. G.; Rayala, S.; Gutti, G.; Sharma, A.; Singh, M.; Kumar, A.; Singh, S. K. Systematic review on role of structure based drug design (SBDD) in the identification of anti-viral leads against SARS-Cov-2. *Current Research in Pharmacology and Drug Discovery* **2021**, *2*, 100026.

(2) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303* (5665), 1813-1818.

(3) De Vivo, M. Bridging quantum mechanics and structure-based drug design. *Front. Biosci* **2011**, *16* (5), 1619-1633.

(4) Schneider, G. Virtual screening: an endless staircase? *Nature Reviews Drug Discovery* **2010**, *9* (4), 273-276.

(5) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2015**, *5* (6), 405-424.

(6) Jorgensen, W. L. Efficient drug lead discovery and optimization. *Accounts of chemical research* **2009**, *42* (6), 724-733.

(7) Jakhar, R.; Dangi, M.; Khichi, A.; Chhillar, A. K. Relevance of molecular docking studies in drug designing. *Current Bioinformatics* **2020**, *15* (4), 270-278.

(8) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry* **2016**, *59* (9), 4035-4061.

(9) Hollingsworth, S. A.; Dror, R. O. Molecular dynamics simulation for all. *Neuron* **2018**, *99* (6), 1129-1143.

(10) Beckstein, O.; Sansom, M. S. A hydrophobic gate in an ion channel: the closed state of the nicotinic acetylcholine receptor. *Physical biology* **2006**, *3* (2), 147.

(11) Marino, K. A.; Shang, Y.; Filizola, M. Insights into the function of opioid receptors from molecular dynamics simulations of available crystal structures. *British journal of pharmacology* **2018**, *175* (14), 2834-2845.

(12) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **2006**, *14* (3), 437-449.

(13) Li, J.; Flick, F.; Verheugd, P.; Carloni, P.; Lüscher, B.; Rossetti, G. Insight into the mechanism of intramolecular inhibition of the catalytic activity of sirtuin 2 (SIRT2). *PLoS One* **2015**, *10* (9), e0139095.

(14) Pantsar, T.; Rissanen, S.; Dauch, D.; Laitinen, T.; Vattulainen, I.; Poso, A. Assessment of mutation probabilities of KRAS G12 missense mutants and their long-timescale dynamics by atomistic molecular simulations and Markov state modeling. *PLoS computational biology* **2018**, *14* (9), e1006458.

(15) Salo-Ahen, O. M.; Alanko, I.; Bhadane, R.; Bonvin, A. M.; Honorato, R. V.; Hossain, S.; Juffer, A. H.; Kabedev, A.; Lahtela-Kakkonen, M.; Larsen, A. S. Molecular

dynamics simulations in drug discovery and pharmaceutical development. *Processes* **2020**, *9* (1), 71.

(16) Borhani, D. W.; Shaw, D. E. The future of molecular dynamics simulations in drug discovery. *Journal of computer-aided molecular design* **2012**, *26*, 15-26.

(17) Coop, A.; MacKerell, A. The future of opioid analgesics. *American Journal of Pharmaceutical Education* **2002**, *66* (2), 153-156.

(18) Healy, J. R.; Bezawada, P.; Shim, J.; Jones, J. W.; Kane, M. A.; MacKerell Jr, A. D.; Coop, A.; Matsumoto, R. R. Synthesis, modeling, and pharmacological evaluation of UMB 425, a mixed μ agonist/δ antagonist opioid analgesic with reduced tolerance liabilities. *ACS chemical neuroscience* **2013**, *4* (9), 1256-1266.

(19) Shaw, D. E.; Grossman, J.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H. 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*, 2014; IEEE: pp 41-53.

(20) Harvey, M. J.; De Fabritiis, G. AceCloud: molecular dynamics simulations in the cloud. ACS Publications: 2015.

(21) Kondratyuk, N.; Nikolskiy, V.; Pavlov, D.; Stegailov, V. GPU-accelerated molecular dynamics: State-of-art software performance and porting from Nvidia CUDA to AMD HIP. *The International Journal of High Performance Computing Applications* **2021**, *35* (4), 312-324.

(22) Turalija, M.; Petrović, M.; Kovačić, B. 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)*, 2022; IEEE: pp 300-306.

(23) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *Journal of chemical theory and computation* **2009**, *5* (6), 1632-1639.

(24) Alam, S.; Varettoa, U. GROMACS on Hybrid CPU-GPU and CPU-MIC Clusters: Preliminary Porting Experiences, Results and Next Steps. **2014**.

(25) Bergdorf, M.; Robinson-Mosher, A.; Guo, X.; Law, K.-H.; Shaw, D. E. Desmond/GPU performance as of April 2021. *DE Shaw Research, Tech. Rep. DESRES/TR–2021-01* **2021**.

(26) Mosalaganti, S.; Obarska-Kosinska, A.; Siggel, M.; Taniguchi, R.; Turoňová, B.; Zimmerli, C. E.; Buczak, K.; Schmidt, F. H.; Margiotta, E.; Mackmull, M.-T. AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* **2022**, *376* (6598), eabm9506.

(27) Nawrocki, G.; Im, W.; Sugita, Y.; Feig, M. Clustering and dynamics of crowded proteins near membranes and their influence on membrane bending. *Proceedings of the National Academy of Sciences* **2019**, *116* (49), 24562-24567.

(28) Pezeshkian, W.; König, M.; Wassenaar, T. A.; Marrink, S. J. Backmapping triangulated surfaces to coarse-grained membrane models. *Nature communications* **2020**, *11* (1), 2296.

(29) Stevens, J. A.; Grünewald, F.; van Tilburg, P. M.; König, M.; Gilbert, B. R.; Brier, T. A.; Thornburg, Z. R.; Luthey-Schulten, Z.; Marrink, S. J. Molecular dynamics simulation of an entire cell. *Frontiers in Chemistry* **2023**, *11*, 1106495.

(30) Luthey-Schulten, Z.; Thornburg, Z. R.; Gilbert, B. R. Integrating cellular and molecular structures and dynamics into whole-cell models. *Current Opinion in Structural Biology* **2022**, *75*, 102392.

(31) Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. Molecular dynamics simulations of large macromolecular complexes. *Current opinion in structural biology* **2015**, *31*, 64-74.

(32) Skånberg, R.; Linares, M.; König, C.; Norman, P.; Jönsson, D.; Hotz, I.; Ynnerman, A. VIA-MD: Visual Interactive Analysis of Molecular Dynamics. In *MolVa@ EuroVis*, 2018; pp 19-27.

(33) Magarkar, A. *MD-Simba*. 2023. https://github.com/aniketsh/MD-SIMBA-Public (accessed.

(34) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: an AI tool for de novo drug design. *Journal of chemical information and modeling* **2020**, *60* (12), 5918-5922.

(35) Glaser, J.; Vermaas, J. V.; Rogers, D. M.; Larkin, J.; LeGrand, S.; Boehm, S.; Baker, M. B.; Scheinberg, A.; Tillack, A. F.; Thavappiragasam, M. High-throughput virtual laboratory for drug discovery using massive datasets. *The International Journal of High Performance Computing Applications* **2021**, *35* (5), 452-468.

(36) Scientific, O. GigaDocking™—Structure Based Virtual Screening of over 1 Billion Molecules Webinar. Aug: 2019.

(37) Rogers, D. M.; Agarwal, R.; Vermaas, J. V.; Smith, M. D.; Rajeshwar, R. T.; Cooper, C.; Sedova, A.; Boehm, S.; Baker, M.; Glaser, J. SARS-CoV2 billion-compound docking. *Scientific Data* **2023**, *10* (1), 173.

(38) Achdout, H.; Aimon, A.; Bar-David, E.; Morris, G. COVID moonshot: open science discovery of SARS-CoV-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *BioRxiv* **2020**.

(39) Boby, M. L.; Fearon, D.; Ferla, M.; Filep, M.; Koekemoer, L.; Robinson, M. C.; Consortium‡, C. M.; Chodera, J. D.; Lee, A. A.; London, N. Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science* **2023**, *382* (6671), eabo7201.

(40) Cournia, Z.; Allen, B.; Sherman, W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *Journal of chemical information and modeling* **2017**, *57* (12), 2911-2937.

(41) Merz Jr, K. M. Limits of free energy computation for protein− ligand interactions. *Journal of chemical theory and computation* **2010**, *6* (5), 1769-1776.

(42) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, 2006; pp 84-es.

(43) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D. OPLS4: Improving force field accuracy on challenging regimes of chemical space. *Journal of chemical theory and computation* **2021**, *17* (7), 4291-4300.

(44) Kythreotou, A.; Siddique, A.; Mauri, F. A.; Bower, M.; Pinato, D. J. PD-L1. *Journal of clinical pathology* **2018**, *71* (3), 189-194.

(45) Yi, M.; Zheng, X.; Niu, M.; Zhu, S.; Ge, H.; Wu, K. Combination strategies with PD-1/PD-L1 blockade: current advances and future directions. *Molecular cancer* **2022**, *21* (1), 28.

(46) Jain, A. N.; Brueckner, A. C.; Jorge, C.; Cleves, A. E.; Khandelwal, P.; Cortes, J. C.; Mueller, L. Complex peptide macrocycle optimization: combining NMR restraints with conformational analysis to guide structure-based and ligand-based design. *Journal of Computer-Aided Molecular Design* **2023**, 1-17.

(47) Wallraven, K.; Holmelin, F. L.; Glas, A.; Hennig, S.; Frolov, A. I.; Grossmann, T. N. Adapting free energy perturbation simulations for large macrocyclic ligands: how to dissect contributions from direct binding and free ligand flexibility. *Chemical science* **2020**, *11* (8), 2269-2276.

(48) Niu, B.; Appleby, T. C.; Wang, R.; Morar, M.; Voight, J.; Villaseñor, A. G.; Clancy, S.; Wise, S.; Belzile, J.-P.; Papalia, G. Protein footprinting and X-ray crystallography reveal the interaction of PD-L1 and a macrocyclic peptide. *Biochemistry* **2019**, *59* (4), 541-551.

(49) Rostkowski, M.; Olsson, M. H.; Søndergaard, C. R.; Jensen, J. H. Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. *BMC structural biology* **2011**, *11*, 1-6.

(50) Miller, M. M.; Mapelli, C.; Allen, M. P.; Bowsher, M. S.; Boy, K. M.; Gillis, E. P.; Langley, D. R.; Mull, E.; POIRIER, M. A.; SANGHVI, N. Macrocyclic inhibitors of the PD-1/PD-L1 and CD80 (B7-1)/PD-L1 protein/protein interactions. Google Patents: 2014.

(51) Jiao, L.; Dong, Q.; Zhai, W.; Zhao, W.; Shi, P.; Wu, Y.; Zhou, X.; Gao, Y. A PD-L1 and VEGFR2 dual targeted peptide and its combination with irradiation for cancer immunotherapy. *Pharmacological Research* **2022**, *182*, 106343.

(52) Watts, K. S.; Dalal, P.; Tebben, A. J.; Cheney, D. L.; Shelley, J. C. Macrocycle conformational sampling with MacroModel. *Journal of chemical information and modeling* **2014**, *54* (10), 2680-2696.

(53) Wu, Y.; Tepper, H. L.; Voth, G. A. Flexible simple point-charge water model with improved liquid-state properties. *The Journal of chemical physics* **2006**, *124* (2).

(54) Berendsen, H. J.; Hayward, S. Collective protein dynamics in relation to function. *Current opinion in structural biology* **2000**, *10* (2), 165-169.

(55) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* **1983**, *4* (2), 187-217.

(56) Lindahl, E.; Hess, B.; Van Der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling* **2001**, *7* (8), 306-317.

(57) Jorgensen, W. L.; Schyman, P. Treatment of halogen bonding in the OPLS-AA force field: application to potent anti-HIV agents. *Journal of chemical theory and computation* **2012**, *8* (10), 3895-3901.

(58) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825-2830.

(59) Klebe, G. Protein-ligand interactions as the basis for drug action. In *Multifaceted Roles of Crystallography in Modern Drug Discovery*, Springer, 2015; pp 83-92.

(60) Kullback, S. Kullback-leibler divergence. 1951.

(61) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J. 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. *Journal of the American Chemical Society* **2015**, *137* (7), 2695-2703.

(62) Schindler, C. E.; Baumann, H.; Blum, A.; Böse, D.; Buchstaller, H.-P.; Burgdorf, L.; Cappel, D.; Chekler, E.; Czodrowski, P.; Dorsch, D. Large-scale assessment of binding free energy calculations in active drug discovery projects. *Journal of Chemical Information and Modeling* **2020**, *60* (11), 5457-5474.

(63) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **2017**, *13* (7), e1005659.

(64) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation* **2008**, *4* (3), 435-447.

(65) Lim, V. T.; Hahn, D. F.; Tresadern, G.; Bayly, C. I.; Mobley, D. L. Benchmark assessment of molecular geometries and energies from small molecule force fields. *F1000Research* **2020**, *9*.

(66) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry* **2011**, *32* (10), 2319-2327.