Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MDFitML: Using machine learning to predict potency from molecular simulations

Xue Zong
OpenEye, Cadence Molecular Sciences
Ahmet Mentes
Bristol Myers Squibb
Barmak Mostofian
OpenEye, Cadence Molecular Sciences
Keith Jones
OpenEye, Cadence Molecular Sciences
David R Stevens
drs73@aya.yale.edu

OpenEye, Cadence Molecular Sciences
Greg Bakken
OpenEye, Cadence Molecular Sciences
Sirish Kaushik Lakkaraju
Bristol Myers Squibb

Research Article

Keywords: MD, ML, protein-ligand interactions, simulation fingerprints, structure-activity relationship

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10529587/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Molecular dynamics (MD) simulations characterize ligand-protein conformational dynamics and can provide a strong structural rationale for measured potency changes accompanying ligand modifications. We present an automated workflow that streamlines not only running MD with a large number of protein-ligand complexes but also identifies critical features that can explain the potency changes within a given set of ligands using machine-learning (ML) models. Simulation fingerprints (SimFP) capture protein-ligand compatibility, including stability of different ligand-pocket interactions that enable easy rationale for rank-ordering of the ligand library during pocket optimization. Here, we present case studies on how ML models that use SimFP as features help delineate structure-activity relationship (SAR) trends and explain potency differences across not only matched-molecular pairs, but also diverse chemotypes of small molecules across protein targets from diverse gene families.

Scientific Contribution

MDFitML introduces simulation fingerprints (SimFP), a molecular descriptor class derived from protein–ligand interaction occupancies over MD trajectories, that enables interpretable ML models in which every prediction maps directly to specific residue-level pocket interactions, a level of structural rationale absent from conventional chemical descriptors and static pose-based features. We demonstrate that linear models trained on SimFP from a single chemotype can accurately rank-order structurally diverse compounds targeting the same pocket, achieving cross-chemotype transferability that remains challenging for both relative free energy methods and descriptor-based ML. The fully automated workflow, validated across eight systems from diverse protein families, provides a scalable and reproducible cheminformatics pipeline for simultaneous potency prediction and mechanistic interpretation in data-limited drug discovery settings.

1. Introduction

Molecular dynamics (MD) simulations provide a physics-based description of protein–ligand interactions that extends beyond static binding poses by explicitly accounting for conformational flexibility, solvent effects, and time dependent interaction behavior and thus occupy a unique position in the methodological landscape of computer-aided drug design[1, 2]. MD trajectories capture the ensemble of conformations and interaction patterns that distinguish closely related analogs, e.g. in congeneric series of small-molecule drug candidates[3]. The time-resolved nature of MD simulations reveals dynamic features, such as the stability of key protein-ligand interactions, that are invisible to static structure-based methods and difficult to encode in conventional molecular descriptors[4]. Although MD simulations are increasingly used in structure based drug discovery, their broader application to quantitative affinity prediction remains limited[5]. Contributing factors include the lack of scalable, automated workflows and the challenge of reducing high dimensional trajectory data to descriptors that are both predictive and chemically interpretable. In contrast, machine learning (ML) approaches have demonstrated utility for potency prediction but typically rely on features derived from either just chemical structures or docked

poses with empirical representations that do not explicitly encode binding dynamics[6]. Consequently, there is a need for approaches that systematically transform MD trajectories into concise, interpretable features suitable for the small and heterogeneous datasets commonly encountered in early discovery projects.

In this study, we extend the MDFit workflow[1] by introducing MDFitML, an automated framework that integrates MD simulations with ML based models that use protein–ligand interaction fingerprints (SimFP) as features. These ML models can be trained against program relevant assay endpoints such as biochemical potencies or binding affinities. In this paper, we will refer to these endpoints as potency wherever appropriate for consistency. By teaching ML models potency differences across a given ligand set using MD interactions, we gain two distinct advantages : 1) ML models can now use interaction fingerprints from MD simulations to directly predict program relevant outcomes such as biochemical potencies; 2) Since the ML models are learning potency differences via MD interactions, there is an unprecedented structural interpretation now brought back to a prediction that is lost when using chemical or other higher dimensional descriptors. Interaction based representations enable the use of simple, regularized linear models that are well suited to data limited regimes and facilitate simultaneous prediction and mechanistic analysis. As such, MDFitML provides a complementary approach to established free energy and pose based ML methods, emphasizing robustness, interpretability, and applicability across diverse protein targets.

2. Methods

2.1 Workflow

MDFitML builds upon previous work[1] by automating the end-to-end process from MD simulations to experimental potency prediction. Specifically, it streamlines the setup of MD simulations, facilitates the analysis of MD trajectories, extracts protein-ligand interactions, and integrates these interactions into machine learning models for potency prediction. The overall workflow is illustrated in Fig. 1, and each component is described in detail in the following sections.

2.2 System preparation

Eight prominent therapeutic targets that cover a variety of protein classes and span diverse cellular functions, such as kinases, regulatory proteins, and transcription factors, were chosen for this study (see Table 1). Protein structures were processed and imported into Orion[7] by the OpenEye Spruce Toolkit[8], building missing loops and capping chain termini. Ligand starting poses for CDK2, MCL-1, BACE1 & Hif-2α were obtained from a previous work by Wang et al.[9] and by Schindler et al.[10] CDK9, Tyk2 and BMS internal target dataset poses were obtained using Glide as previously described[1].

2.3 MD simulations

All MD simulations were performed at all-atom resolution using the MD Orion floes (OpenMM[11] engine). The protein-ligand complexes were solvated in an aqueous solution at 0.15 M NaCl, producing simulation system sizes that ranged from ~ 20,000 to ~ 77,000 atoms.

The AMBER14SB force field[12] was used for the proteins and OpenFF 2.0.0[13] was applied to all ligand molecules with the TIP3P water model[14]. Long-range electrostatic interactions were calculated using the Particle-Mesh Ewald summation[15, 16]. Short-range Coulomb and van der Waals (vdW) interactions were cut off at 10 Å. Bonds to hydrogen atoms were constrained using the Constant Constraint Matrix Approximation (CCMA)[17] and the SETTLE algorithm[18]. The solvated protein-ligand systems were first energy-minimized until convergence using the L-BFGS method[19], followed by an equilibration simulation phase of 0.25 ns, during which positional restraints on heavy atoms were gradually removed from 2 to 0 kcal/mol/Å^2, the simulation time step was increased from 1 fs to 4 fs, enabled by hydrogen mass repartitioning[20], and the simulation ensemble was switched from NVT to NPT. Langevin dynamics were applied to control the temperature at 300 K with a damping coefficient of 5/ps[21], using the LFMiddle discretization integrator[22, 23], and the isotropic Monte Carlo barostat was applied to keep the pressure at 1 atm. All MD simulations were run for 100 ns and saved at 0.1ns intervals. The first 10 ns (100 simulation frames) of each simulation were used for system equilibration and thus ignored for the subsequent analysis. To increase the independent sampling and thus the number of instances for the regression analysis, all 8 proteins were simulated with each of the ligands from their corresponding collection in triplicate.

2.4 SimFP

Protein-ligand interactions were generated using OEChem TK[24], capturing interactions such as contact, hydrogen bond, salt-bridge, stacking (T and Pi), among others (see Supplementary Information for details). In addition, we incorporated water-mediated protein-ligand interactions, defined when a water molecule simultaneously makes a hydrogen bond with the protein and with the ligand. All interactions were determined based on pre-defined distances and angles derived from previous literature[25–28]. To utilize interactions as features for ML model development, we define the interaction feature descriptors as follows. For static poses, each interaction descriptor is defined by the combination of interaction type, protein residue name and the residue number. Each descriptor represents the frequency with which a particular interaction occurs and can exceed unity when multiple atoms from a single protein residue interact with multiple ligand atoms. For MD simulation data, we calculate the interaction occupancy rather than simple interaction counts. Interaction occupancy is computed by summing interaction counts across all MD frames and dividing by the total number of frames after removing the first 10 ns of data. This workflow systematically extracts and tabulates all protein-ligand interactions, quantifying their stability as the percentage of simulation time during which each interaction is observed.

2.5 ML model development

After extracting SimFP, we used them as features to train ML models for potency prediction and to provide insights into key potency determinants. Given the limited dataset size, we focused on simple, interpretable ML models that not only enable potency prediction but also facilitate identification of critical interactions. To maximize the utility of our MD dataset, we treated each of the three replicate MD simulations per compound as independent training instances during model development. This approach effectively augmented our training data while preserving the statistical independence required for robust model learning. For final predictions, we employed an ensemble approach by averaging the predictions from all three MD replicates for each compound, thereby reducing prediction variance and improving overall model reliability. To address the curse of dimensionality (i.e., a large number of interaction features relative to the number of compounds), we designed a two-step modeling strategy.

First, a Lasso regression model was applied for feature selection, identifying important interactions based on regression coefficients. Next, a simple linear regression model was trained using the selected features to predict potency. All models were implemented in scikit-learn[29]. For the Lasso model, hyperparameters were optimized using nested leave-one-molecule-out cross-validation (LOMO-CV), following prior literature[1]. Feature matrices were standardized using MinMaxScaler from scikit-learn[29] to ensure that all input variables were on a comparable scale. The response variable was represented in either pIC50 or DG (binding free energy). Feature importance was computed as the average coefficient values across cross-validation folds. CoD (coefficients of determination) and RMSE (root-mean-squared-error) were calculated to evaluate the model performance and listed in Table 1.

During exploratory analysis, we also tested more complex models such as Random Forest. However, learning curve analysis revealed substantial overfitting with these approaches (see Supplementary Information) due to the limited size of datasets, and thus we adopted the simpler workflow described above.

3. Results and Discussion

Table 1
ML model performance across investigated systems. RMSE in kcal/mol for datasets with DG response. 95% confidence intervals were calculated via bootstrapping and details are provided in Table S1.

Protein target (PDB ID)	Docked Poses + ML				MDFitML (MD + ML)		
	No. ligands	No. interactions	CoD	RMSE	No. interactions	CoD	RMSE
CDK2 (1H1Q)[30]	16	27	0.57	0.77	136	0.90	0.38
CDK9 (7NWK)[31]	36	53	0.68	0.63	227	0.48	0.81
MCL-1(4HW3)[32]	42	25	0.10	1.00	166	0.70	0.58
BACE1 (4DJW)[33]	36	34	0.30	0.66	162	0.77	0.38
Hif-2α (5TBM)[34]	42	31	0.24	0.98	208	0.34	0.91
Tyk2[35] – nicotinamide chemotype	66	53	0.46	0.58	184	0.5	0.6
BMS internal target-Chemotype A	112	70	0.34	0.77	289	0.42	0.76
BMS internal target-Chemotype B	147	65	0.29	0.62	296	0.65	0.43

Datasets for five different targets CDK2, CDK9, MCL-1, BACE1 and Hif-2α were obtained from a previous work that benchmarked free-energy based methods performance against these sets[36]. In this work, we assessed the ability of the linear regression model built using the top 10 interactions to correctly explain the potency differences among these datasets. CoD & RMSE for the MDFitML quantify these capabilities of the linear regression model that is built purely using the top 10 SimFP interactions identified using the Lasso model. For Tyk2 and the BMS internal target, we extended this to probe the transferability of a regression model thus built using SimFP to go beyond the training set and the chemotype used and rank order new diverse ligands. Can a regression model trained using SimFP and potency data from one chemotype interactions with the pocket, correctly predict the potency of a chemically diverse molecule, so long as it binds to the same pocket? Tyk2 JH2 pseudodomain[37] was previously identified to offer a selective binding pocket and several different inhibitors across multiple chemotypes have been reported to engage with this pocket. In this study, we picked two chemotypes : nicotinamides[35] and imidazopyridazines (IZP)[38]. A regression model was first built using nicotinamides as a training set and then the transferability of that regression model was tested using compounds from the imidazopyridazine chemotype (see below).

A similar exercise was run with another BMS internal target. Two structurally distinct chemotypes: Chemotype A, comprising a biaryl scaffold series, and Chemotype B, based on an acetylene scaffold

series were used. For both chemotype series within the BMS internal target dataset, heterogeneous R-group substitution patterns, rather than discrete, single R-group variations, were incorporated into the MD simulations, allowing for a more comprehensive sampling of the chemical space and substituent conformational diversity. Like the Tyk2, two separate MDFitML models were built and ML model built using data from chemotype A was capable of rank-ordering chemotype B. As shown in Table 1 and discussed in the sections 3.2.2 below these regression models are capable of correctly predicting potencies of any molecule, so long as they occupy and interact with the similar binding pocket and we can characterize these molecules as SimFPs that the regression model knows about. This therefore opens up a wealth of possibilities on how MD data can come back to enable design triaging even as drug discovery projects pivot several times from a known chemotype for multiple reasons beyond improving potency[39].

We also compared the predictive performance of ML models built from static docked poses versus MDFitML. The MDFitML approach consistently generated substantially richer feature spaces, capturing more than 100 protein-ligand interactions per target system compared to only 25–53 interactions from docked poses. This increased feature dimensionality reflects the dynamic nature of protein-ligand binding captured through MD simulations, where transient interactions are better sampled over the simulation trajectory. For most target proteins (except for CDK9), MDFitML models demonstrated better predictive performance. CDK2 showed marked improvement (CoD from 0.57 to 0.90), MCL-1 exhibited dramatic enhancement (CoD from 0.10 to 0.70), and BACE1 also displayed substantial gains (CoD from 0.30 to 0.77). These improvements can be attributed to MD simulations accounting for protein flexibility, solvent effects, and the dynamic interplay between ligand and protein, thereby capturing more biologically relevant binding features that corelate with experimental affinities. However, the model performance also varied across targets. CDK9 showed reduced correlation with MDFitML, while Hif-2α exhibited comparable RMSE values between the two approaches despite generating 208 interactions in MDFitML. These exceptions suggest that the benefits of MD-derived features may be target-dependent, potentially influenced by factors such as binding site flexibility, ligand strain[1] etc for certain scaffolds.

The dataset sizes in this study ranged from 16 ligands (CDK2) to 147 ligands (BMS internal dataset), representing relatively small training sets typical of early-stage drug discovery projects. Despite the limited data, MDFitML models were able to correctly learn which interactions explain the potency variations within a dataset. This reflects the ability of SimFP and ML models that learn from them to correctly account for dynamic binding processes, including transient interactions due to protein conformational changes, that are absent in static structures to explain potency changes.

3.1 ML model development

To examine the robustness and design choices underlying MDFitML, we conducted a series of methodological experiments focused on the five public target systems (CDK2, CDK9, MCL-1, BACE1 and Hif-2α). These systems were selected for this exploration because their experimental data, binding poses, and structural context are publicly available, enabling transparent benchmarking and

reproducibility. In the following subsections, we systematically evaluate four key modeling decisions: the choice between full Lasso regression and a reduced linear model using the top interactions, the statistical validity of the learned models via Y-scrambling, the effect of interaction occupancy filtering on feature dimensionality and predictive accuracy, and the sensitivity of model performance to the number of included interactions.

3.1.1 Lasso regression vs Linear model using top 10 interactions. As described in the Methods section, we implemented a two-step modeling strategy to address the challenge of building predictive models in a high-dimensional, low-sample-size setting. In the first step, we applied Lasso regression on the complete feature set, leveraging its L1 regularization to identify the most informative predictors. The Lasso model's built-in feature selection mechanism provided a data-driven ranking of feature importance while managing multi-collinearity inherent in interaction terms. In the second step, we selected the top 10 interactions based on their regression coefficients and trained a standard linear regression model using only this reduced feature set.

The performance comparison between the two approaches revealed substantial improvements when using the reduced linear model (see Table 2). Across all five public protein targets, the linear models trained on the top 10 Lasso-selected features demonstrated superior predictive accuracy compared to the full Lasso models. For instance, CDK2 showed a reduction in RMSE from 1.00 to 0.38, while the coefficient of determination improved from 0.28 to 0.90. Similar patterns emerged for the other targets, with RMSE reductions ranging from 35% to 58% and corresponding improvements in CoD.

The superior performance of the reduced linear model can be attributed to several factors. First, by distilling the feature space from over 200 to 10 interactions, we dramatically improved the sample-to-feature ratio from 0.4 to 10, moving from a severely overdetermined system to one where robust parameter estimation is feasible. Second, while Lasso regularization helps prevent overfitting, the penalty term still introduces bias that can limit predictive accuracy. Once the most relevant features are identified, removing this constraint allows the linear model to capture the true relationships more precisely. Third, the Lasso model effectively handles the feature selection problem by identifying which predictors matter, but the final linear model benefits from the increased statistical power available when estimating fewer parameters with the same sample size. This two-step approach thus combined the feature selection strength of Lasso with the unbiased estimation properties of ordinary least squared in an appropriately dimensioned feature space.

Table 2

Performance comparison of Lasso and Linear regression models across five public protein targets. Lasso models were trained on the complete dataset with all interaction features, while Linear regression models were trained using only the top 10 features selected based on regression coefficients. Performance metrics include coefficient of determination (CoD) and root mean squared error (RMSE). 95% confidence intervals from bootstrapping are included in Table S2.

Protein Target	Lasso		Linear with Top 10 Interactions	
	CoD	RMSE	CoD	RMSE
CDK2	0.28	1.00	0.90	0.38
CDK9	-0.37	1.32	0.48	0.81
BACE1	-3.16	1.63	0.77	0.38
Hif-2α	-0.57	1.41	0.34	0.91
MCL-1	0.34	0.85	0.70	0.58

3.1.2 Y-scrambling experiments. To validate the robustness of our predictive models and eliminate the possibility of spurious correlations arising from limited dataset sizes combined with high-dimensional feature spaces, we performed Y-scrambling experiments on all five public target systems. Y-scrambling (also known as Y-randomization), is a standard validation technique that involves randomly permuting the target variable while keeping the feature matrix unchanged, thereby destroying any genuine relationship between predictors and outcomes while preserving the statistical structure of the predictor variables[40]. For each system, we generated 100 independent scrambled datasets and retrained models on each permuted dataset to establish a null distribution of performance metrics. The underlying principle is that if the original models capture meaningful structure-activity relationships rather than chance correlations, they should significantly outperform models trained on scrambled data where no true predictive signal exists. To ensure robust statistical assessment, we implemented outlier removal. If any scrambled model produced anomalous bad predictions with CoD less than − 1 (typically due to random chance producing an unusually unfavorable split), that iteration was excluded before calculating summary statistics. This conservative approach ensures that the reported scrambling results represent the best possible performance achievable by chance alone. Both CoD and RMSE were calculated to compare with MDFitML results.

Table 3
Y-scrambling experiment results on five public systems. 95% confidence interval details are included in Table S3.

System	MDFitML Results		Scramble Experiments Results	
	CoD	RMSE	CoD	RMSE
CDK2	0.90	0.38	0.32	0.93
CDK9	0.48	0.81	0.27	0.95
BACE1	0.77	0.38	0.21	0.70
Hif-2α	0.34	0.91	0.21	0.99
MCL-1	0.70	0.58	0.17	0.95

The Y-scrambling results shown in Table 3 conclusively demonstrate that MDFitML models do not simply capture spurious correlations. All systems showed substantial performance degradation upon scrambling. CDK2 exhibited a 64% reduction with CoD dropping from 0.90 to 0.32, BACE1 showed a 73% decline (CoD from 0.77 to 0.21), while CDK9 and Hif-2α displayed reductions of 44% and 38%, respectively. These dramatic decreases in predictive power were accompanied by corresponding increases in prediction errors, with RMSE values rising 1.8- to 2.3-fold across all targets. Critically, the 95% confidence intervals (see Table S3) for original and scrambled models showed minimal or no overlap for most systems, confirming statistical significance of the performance differences. This consistent pattern of substantial performance degradation upon scrambling provides robust statistical evidence that the MD-derived interaction occupancy features encode genuine physicochemical determinants of binding affinity, validating the MDFitML approach for predictive modeling even with relatively small training datasets.

3.1.3 Effects of varying interaction occupancy threshold. Given the high-dimensional nature of our interaction occupancy features for MD datasets, where the number of features substantially exceeded the number of training compounds, we investigated whether applying occupancy-based filtering could improve model performance and interpretability. We systematically explored different occupancy cutoff thresholds to eliminate interactions that appeared infrequently across the MD trajectories. Specifically, an interaction was retained only if it appeared in at least cutoff threshold * N frames, where N represents the total number of MD frames analyzed. We tested cutoffs of 10% and 20% to evaluate their impact on feature dimensionality and predictive accuracy.

As expected, applying stricter cutoff thresholds substantially reduced the feature space. Across all systems, applying 10% cutoff threshold reduced the number of interactions by approximately 50–65% (e.g. from 227 to 99 interactions for CDK9, from 166 to 73 for BACE1), while the 20% cutoff threshold further decreased features by an additional 20–30%. This dramatic reduction reveals that most protein-ligand interactions observed during MD simulations are transient in nature, persisting for less than 10% of the trajectory. These transient interactions may represent conformational fluctuations, solvent-

mediated contacts, or exploratory binding modes that do not significantly contribute to the overall binding. However, they could also capture dynamic aspects of binding that are relevant in certain contexts.

The impact of these cutoff thresholds proved to be system-dependent, with no single threshold consistently improving predictive performance across all examined protein-ligand systems. While some systems (Hif-2α and MCL-1) showed modest improvements at intermediate cutoffs, others exhibited degraded performance or negligible changes. These results suggest that the relationship between interaction frequency and binding affinity is complex and system-specific, and that even relatively transient interactions may contain valuable information for certain protein-ligand complexes. In some cases, removing transient interactions reduced noise and improved generalization, while in others, these dynamic features appeared to contain valuable predictive information. Given this system-dependent behavior and the absence of a universally optimal threshold, we retained all interaction features in our final models and instead relied on regularization techniques to manage the high-dimensional feature space effectively.

Table 4
Summary of model performance when applying different cutoff thresholds across five public systems. Performance metrics include coefficient of determination (CoD) and root mean squared error (RMSE). 95% confidence interval from bootstrapping are included in Table S4.

Systems	Cutoff	No. Interactions	CoD	RMSE
CDK2	0	136	0.90	0.38
	10%	61	0.86	0.44
	20%	48	0.78	0.54
CDK9	0	227	0.48	0.81
	10%	99	0.23	0.98
	20%	76	0.42	0.86
BACE1	0	166	0.77	0.38
	10%	73	0.56	0.53
	20%	57	0.61	0.50
Hif-2α	0	162	0.34	0.91
	10%	59	0.45	0.83
	20%	50	0.59	0.73
MCL-1	0	208	0.70	0.58
	10%	86	0.53	0.72
	20%	62	0.73	0.54

3.1.4. Effects of number of included interactions. For our primary analysis, we selected the top 10 interactions to build the Linear regression model. Here we systematically evaluate model performance across varying feature set sizes (5, 10, 20, 30, 40, and 50 interactions) to assess the robustness of this choice.

The results (see Fig. 2) demonstrate that model performance varies considerably with the number of included interactions, though the specific trends differ across protein targets. For CDK9 and MCL-1, the coefficient of determination remains relatively stable across all tested interaction counts, with values considerably above 0.5–0.9, indicating robust predictive performance. However, for BACE1 and CDK2, we observe substantial performance degradation when incorporating 50 interactions, with CoD values dropping precipitously to negative values and RMSE increasing dramatically (from ~ 0.3–0.5 to 1.0–3.0). This pattern is characteristic of overfitting, where the model begins to capture noise rather than genuine signal when the feature-to-sample ratio becomes unfavorable. Notably, all five targets achieve their

optimal or near-optimal performance within 10–30 interaction range, with minimal improvement or even deterioration beyond this threshold.

These findings validate our selection of 10 interactions for the primary models, demonstrating that this choice strikes an appropriate balance between predictive accuracy and model interpretability. The dramatic performance collapse observed for some targets at higher interaction counts underscores the importance of feature selection in high-dimensional, low-sample regimes and confirms that our Lasso-based feature selection step successfully identifies the most informative interactions while avoiding overfitting. The consistency of this pattern across multiple diverse protein targets suggests that our two-step modeling framework is broadly applicable in data-limited scenarios.

3.2.1 MDFitML explains potency differences of matched pairs. As noted in Table 1, MDFitML is capable of explaining potency differences among a ligand set using binding pocket interactions across 7 different protein targets. Here, we take a deeper dive to illustrate the interpretability of this across matched molecular pairs targeting Cyclin dependent kinase CDK2, a member of the CDK family and is a validated target for tumor-selective therapeutic strategies[41, 42]. Wang et al[41] showed that dG calculated using Schrodinger's free energy perturbation with high quality co-crystal structures as starting poses have strong correlation with the measured Ki for a congeneric set of ligands. We took this dataset of 16 co-crystal structures and put them through our MDFitML workflow. As shown in Table 1, linear regression model's assessment of dG that is built using the top 10 identified SimFP, correlates quite well with that derived from measured Ki (Fig. 3A). Thus, our approach is capable of explaining the potency differences among these congeneric series of ligands that share the overall binding pose to the hinge. Our Lasso model identified that beyond the hinge interactions, contact interaction with Lys89, is found to be quite important in explaining differences in potencies among these ligands. (Fig. 3B). In contrast to weak binders, all tight binders carry a sulfonamide group that picks up interactions with this Lys89. Interestingly, the exact arrangement of this functional group in the CDK2 binding site is critical for binding affinity. For instance, **1h1s** and **1oiu** are identical compounds except for the sulfonamide being para with respect to the amino-phenyl in the tight binder compared to being meta in in 1oiu. MD with h1s reveals a stronger interaction with Lys89 that can be tied back to its improved potency to this pocket compared to 1oiu with little to no Lys89 interaction. Similarly, lack of this sulfonamide in 17 also correlates with drop in its potency comparable to 1oiu (Fig. 3C-H).

3.2.2 MDFitML regression model can be transferred to other chemotypes. Our tests with five diverse public datasets confirmed that MDFitML is capable of explaining potency differences within a ligand set using SimFP. We next extended this to ask, can we deploy the regression model to prospectively estimate potency of a ligand that is not a part of the training set? And since the ligand is being featurized using only SimFP, can we estimate potencies of ligands from significantly diverse chemotypes compared to the training ligand set? It has been reported that both classical ML models[43–45] as well as relative free energy methods[36, 46, 47] exhibit increased uncertainty of predictions and therefore require much larger training sets/longer simulation times to tackle chemically diverse query molecules. Our hypothesis is that as long as the query molecule occupies the same binding pocket and can be featurized using

similar SimFP as the training set, our MDFitML regression model should be capable of estimating its potency, no matter how chemically diverse it is compared the training set.

To probe this, we picked two diverse ligand sets targeting the JH2 pocket of Tyrosine kinase 2 (Tyk2), a member of the JAK kinase family that regulates IL-23/IL12 pathway mediated signal transduction[48]. BMS previously reported two highly selective Tyk2 inhibitor chemotypes nicotinamides[35] and imidazopyridazines[38] that target the JH2 pocket. So, to test the transferability of a MDFitML regression model, we did the following: We first ran MD with 66 compounds from the nicotinamide series. High quality co-crystal structures[35] of the nicotinamides bound to JH2 were used as starting poses when available. Rest of the compounds were docked into this pocket using Glide[49] to preserve the overall binding pose to the hinge by the nicotinamides using docking constraints as previously described[1]. From these simulations, as before, top 10 SimFP were identified. A linear regression model was developed using these 10 SimFP to explain differences in measured potencies.

Next, 11 compounds from the IZP chemotypes were put through MD. Akin to the nicotinamides, high quality co-crystal structures were used when available[38]. The rest of the compounds within the series were docked following similar docking constrains to preserve the overall chemotype interactions with the hinge region. From these MD simulations, SimFPs were constructed. Now, instead of building a new regression model, we re-used the nicotinamides regression model to estimate the potencies of the IZP series. As shown in Fig. 4, MDFitML regression model built using nicotinamides is quite capable of explaining the potency differences even within the IZPs. We note that the inverse is not true. Regression model built using IZPs MD however, could not accurately predict the potencies of the nicotinamides (R^2 =0.2, RMSE = 1.4), hinting that MDFitML's ability to correctly identify the relevant SimFP that can explain potency differences might be dependent on the size of the training set. Nevertheless, it was interesting to note that a robust regression model built using nicotinamides was capable of explaining potency differences beyond the training set, thereby highlighting the transferability of this approach in assessing a larger chemical diversity of ligands.

Next, a similar analysis was also performed with an internal BMS target. As with Tyk2, two chemically diverse chemotypes were identified to bind and inhibit this internal protein target. Chemotype A consisted of a biaryl scaffold while chemotype B conserved an acetylene scaffold. MDFit was run with 112 chemotype A compounds, top 10 pocket interactions were characterized and a linear regression model was developed using this set. A second independent MDFitML was also run with 147 chemotype B compounds and a separate regression model was developed with this set. Independently, MDFitML is capable of explaining the potency differences within each of these chemotypes using these separate regression models (Table 1). Interestingly, like the Tyk2 dataset, cross-chemotype predictive capability was also apparent here. Regression model built exclusively using 112 chemotype A (biaryls) compounds, was able to rank-order with reasonable accuracy (R^2 = 0.31, RMSE = 0.81) the 147 chemotype B (acetylene) compounds. This suggests that despite chemical divergence, so long as the compounds

bind to the same pocket that was used in the regression model training, the model is capable of predicting and classifying those with strong vs. poor potency.

4. Conclusions

We presented an automated workflow that streamlines not only running MD with large number of protein-ligand complexes but also identifies critical features that can explain the potency changes within a given set of ligands using ML models. MDFitML produces compiled SimFPs for users to decipher critical protein-ligand interactions and the ML model helps a) rank-order ligands based on pocket compatibility, b) explain potency differences among a given ligand set, c) a generic linear regression model that can be used to predict potency of any ligand that will bind to the pocket from the training set. Since the ML models are learning potency differences via MD interactions, there is an unprecedented structural interpretation now brought back to a prediction that is lost when using chemical or other higher dimensional descriptors.

Across 5 public and 2 internal datasets we demonstrate that MDFitML is quite robust in explaining potency differences via structural rationale among congeneric ligands. We also showed that the regression model built using the top 10 SimFP for a given binding pocket is capable of predicting potency correctly for ligands that are significantly diverse compared to the training set.

A fundamental challenge shared by both free-energy and ML models that use high-dimensional descriptors for property prediction is that their reliability diminishes as query molecules diverge from well-characterized chemical space. In relative binding free energy methods, prediction accuracy is governed by phase space overlap between alchemical end states; as the chemical dissimilarity between ligand pairs increases—through scaffold hops, ring system transformations, or net charge modifications —this overlap erodes, leading to poor convergence and unreliable free energy estimates[47, 50, 51]. Perturbation network design tools such as LOMAP[52] explicitly enforce similarity thresholds between connected ligand pairs, effectively restricting RBFE calculations to congeneric series. Although strategies including chemical intermediates[53, 54], separated-topology approaches[55], and enhanced sampling techniques[56] have been developed to extend the reach of these methods, they do so at substantially increased computational cost and methodological complexity. An analogous limitation constrains data-driven approaches. For message-passing neural networks such as Chemprop's D-MPNN[57], the most important determinant of prediction accuracy is the structural similarity of the query molecule to the training set, often outweighing the choice of algorithm itself[58]. Taken together, these observations highlight a shared vulnerability: both approaches are fundamentally interpolative, performing well within well-sampled regions of chemical space but exhibiting increasing uncertainty as predictions extend toward structurally novel territory. While uncertainty quantification methods—including cycle closure analysis for FEP networks[59, 60] and deep ensembles, evidential deep learning, or conformal prediction for ML models[61–63] - can help flag unreliable predictions, they quantify rather than resolve the underlying limitation. This convergence of failure modes underscores the validity of MDFitML

as an attractive alternative to reliably traverse large chemical distances while preserving interpretability via pocket interactions. As demonstrated with the Tyk2 JH2 pocket and the BMS internal target, so long as the query ligand occupies the same binding pocket and can be featurized using SimFP that are a part of the training set, prediction accuracy seems reliably good using MDFitML.

The entire workflow is fully enabled to work on the Orion[64] compute infrastructure that is built for scale. This further drives the ability to increase the throughput of the number of ligands that can be assessed using MDFitML. We continue to be excited about the future of applying these automated MD approaches in enabling drug discovery projects going forward.

Abbreviations

BACE1, beta-site amyloid precursor protein cleaving enzyme 1; CDK, cyclin-dependent kinase; CCMA, Constant Constraint Matrix Approximation; CoD, coefficient of determination; FEP, free energy perturbation; HIF-2α, hypoxia-inducible factor 2 alpha; IZP, imidazopyridazine; JAK, Janus kinase; LASSO, least absolute shrinkage and selection operator; LOMO-CV, leave-one-molecule-out cross-validation; LOMAP, lead optimization mapper; MCL-1, myeloid cell leukemia 1; Tyk2, tyrosine kinase 2.

Declarations

Availability of Data and Materials

All publicly available datasets for CDK2, CDK9, MCL-1, BACE1, and Hif-2α can be found from previous works (reference 9 and 10). The BMS internal target datasets are proprietary and not publicly available. Protein structures were prepared using Spruce TK (reference 8), and protein–ligand interaction fingerprints were computed using OEChem TK (reference 13). MD simulations were carried out with OpenMM (reference 11) through the Orion platform (reference 7, 43), using the AMBER14SB protein force field, the Open Force Field 2.0.0 (Sage) small-molecule force field (reference 12), and TIP3P water model, as detailed in the Methods section. Interaction occupancy matrices (SimFP) derived from MD trajectories are provided in the Supplemental Information. ML models were built using the Lasso and Linear Regression classes from scikit-learn (reference 15), with feature scaling via MinMaxScaler, as described in the Methods section. MDFitML floes will be available on the academic stack of Orion, which is also available upon request.

Supplemental Information

Complete list of protein-ligand interaction types considered in this work; visualization of example interactions (Fig S1); learning curve analysis of Random Forest models demonstrating overfitting in the low-sample, high-dimensional data regime (Fig S2); ML model performance with 95% confidence intervals from bootstrapping across all investigated systems (Table S1); performance of Lasso versus Linear Regression models with confidence intervals (Table S2); Y-scrambling experiment results with confidence intervals (Table S3); model performance under varying interaction occupancy cutoff

thresholds with confidence intervals (Table S4). CSV files detailing the interaction hints for each of the five public datasets.

Corresponding Author

David R Stevens – OpenEye, Cadence Molecular Sciences, Santa Fe, NM 87508; http://orcid.org/0000-0002-7679-7955; Email: drs73@aya.yale.edu

Authors' Contributions

X.Z.: Conceptualization, Methodology, Data Curation, Software, Formal Analysis, Writing. A.M.: Conceptualization, Data Curation, Formal Analysis, Methodology, Writing. B.M.: Data Curation, Formal Analysis, Methodology, Writing. K.J.: Software. D.R.S.: Conceptualization, Supervision, Writing. G.B.: Conceptualization, Supervision. S.K.L.: Conceptualization, Methodology, Data Curation, Formal Analysis, Writing

Funding

The authors have no outside funding to declare.

Completing interests

The authors declare no competing interests.

Acknowledgements

The authors wish to thank Shree Soundarya SV & Veerabahu Shanmugasundaram at Bristol Myers Squibb for helpful discussions around the development and implementation of MDFit-ML. The authors also thank Geoff Skillman and Brenda Montalvo Ortiz from OpenEye, Cadence Molecular Sciences for their support on this project.

References

1. Brueckner AC, Shields B, Kirubakaran P, Suponya A, Panda M, Posy SL, Johnson S, Lakkaraju SK (2024) MDFit: automated molecular simulations workflow enables high throughput assessment of ligands-protein dynamics. J Comput Aided Mol Des 38(1):24
2. 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
3. Malvacio I, Buonfiglio R, D'Atanasio N, Serra G, Bosin A, Di Giorgio FP, Ruggerone P, Ombrato R, Vargiu (2019) A. V. Molecular basis for the different interactions of congeneric substrates with the polyspecific transporter AcrB. Biochim et Biophys Acta (BBA)-Biomembranes 1861(7):1397–1408
4. Alonso H, Bliznyuk AA, Gready JE (2006) Combining docking and molecular dynamic simulations in drug design. Med Res Rev 26(5):531–568

5. Husic BE, Pande VS (2018) Markov state models: From an art to a science. J Am Chem Soc 140(7):2386–2396
6. Fischer Y, Southiratn T, Triki D, Cedeno R (2025) Deep Learning vs Classical Methods in Potency and ADME Prediction: Insights from a Computational Blind Challenge. J Chem Inf Model 65(24):13115–13131
7. Orion (2024). *1*; http://www.eyesopen.com (accessed
8. Spruce TK, *1.7.0.1*; (accessed
9. 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
10. Schindler CE, Baumann H, Blum A, Bose 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
11. 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
12. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput 11(8):3696–3713
13. Boothroyd S, Behara PK, Madin OC, Hahn DF, Jang H, Gapsys V, Wagner JR, Horton JT, Dotson DL, Thompson MW (2023) Development and benchmarking of open force field 2.0. 0: the sage small molecule force field. J Chem Theory Comput 19(11):3251–3275
14. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79(2):926–935
15. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. J Chem Phys 98(12):10089–10092
16. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen L (1995) G. A smooth particle mesh Ewald method. J Chem Phys 103(19):8577–8593
17. Eastman P, Pande VS (2010) Constant constraint matrix approximation: a robust, parallelizable constraint method for molecular simulations. J Chem Theory Comput 6(2):434–437
18. Miyamoto S, Kollman PA, Settle (1992) An analytical version of the SHAKE and RATTLE algorithm for rigid water models. J Comput Chem 13(8):952–962
19. Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. Math Program 45(1):503–528
20. Hopkins CW, Le Grand S, Walker RC, Roitberg AE (2015) Long-time-step molecular dynamics through hydrogen mass repartitioning. J Chem Theory Comput 11(4):1864–1874

21. Pastor RW, Brooks BR, Szabo A (1988) An analysis of the accuracy of Langevin and molecular dynamics algorithms. Mol Phys 65(6):1409–1419
22. Zhang Z, Liu X, Yan K, Tuckerman ME, Liu J (2019) Unified efficient thermostat scheme for the canonical ensemble with holonomic or isokinetic constraints via molecular dynamics. J Phys Chem A 123(28):6056–6079
23. Åqvist J, Wennerström P, Nervall M, Bjelic S, Brandsdal BO (2004) Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. Chem Phys Lett 384(4–6):288–294
24. *OEChem TK*; (accessed
25. Kumar S, Nussinov R (2002) Relationship between ion pair geometries and electrostatic strengths in proteins. Biophys J 83(3):1595–1612
26. Cavallo G, Metrangolo P, Milani R, Pilati T, Priimagi A, Resnati G, Terraneo G (2016) The halogen bond. Chem Rev 116(4):2478–2601
27. Bissantz C, Kuhn B, Stahl M (2010) A medicinal chemist's guide to molecular interactions. J Med Chem 53(14):5061–5084
28. Marcou G, Rognan D (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inf Model 47(1):195–207
29. 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
30. Davies TG, Bentley J, Arris CE, Boyle FT, Curtin NJ, Endicott JA, Gibson AE, Golding BT, Griffin RJ, Hardcastle IR (2002) Structure-based design of a potent purine-based cyclin-dependent kinase inhibitor. Nat Struct Biol 9(10):745–749
31. 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
32. Friberg A, Vigil D, Zhao B, Daniels RN, Burke JP, Garcia-Barrantes PM, Camper D, Chauder BA, Lee T, Olejniczak ET (2013) Discovery of potent myeloid cell leukemia 1 (Mcl-1) inhibitors using fragment-based methods and structure-based design. J Med Chem 56(1):15–30
33. Cumming JN, Smith EM, Wang L, Misiaszek J, Durkin J, Pan J, Iserloh U, Wu Y, Zhu Z, Strickland C (2012) Structure based design of iminohydantoin BACE1 inhibitors: identification of an orally available, centrally active BACE1 inhibitor. Bioorg Med Chem Lett 22(7):2444–2449
34. Wallace EM, Rizzi JP, Han G, Wehn PM, Cao Z, Du X, Cheng T, Czerwinski RM, Dixon DD, Goggin BS (2016) A small-molecule antagonist of HIF2α is efficacious in preclinical models of renal cell carcinoma. Cancer Res 76(18):5491–5500
35. Wrobleski ST, Moslin R, Lin S, Zhang Y, Spergel S, Kempson J, Tokarski JS, Strnad J, Zupa-Fernandez A, Cheng L (2019) Highly selective inhibition of tyrosine kinase 2 (TYK2) for the treatment of autoimmune diseases: discovery of the allosteric inhibitor BMS-986165. J Med Chem 62(20):8973–8995

36. Ross GA, Lu C, Scarabelli G, Albanese SK, Houang E, Abel R, Harder ED, Wang L (2023) The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. Commun Chem 6(1):222
37. Tokarski JS, Zupa-Fernandez A, Tredup JA, Pike K, Chang C, Xie D, Cheng L, Pedicord D, Muckelbauer J, Johnson SR et al (2015) Tyrosine Kinase 2-mediated Signal Transduction in T Lymphocytes Is Blocked by Pharmacological Stabilization of Its Pseudokinase Domain. J Biol Chem 290(17):11061–11074. 10.1074/jbc.M114.619502 From NLM Medline
38. Liu C, Lin J, Moslin R, Tokarski JS, Muckelbauer J, Chang C, Tredup J, Xie D, Park H, Li P (2019) Identification of imidazo [1, 2-b] pyridazine derivatives as potent, selective, and orally active Tyk2 JH2 inhibitors. ACS Med Chem Lett 10(3):383–388
39. Tibbitts J, Canter D, Graff R, Smith A, Khawli LA (2016) Key factors influencing ADME properties of therapeutic proteins: A need for ADME characterization in drug discovery and development. In *MAbs*, ; Taylor & Francis: Vol. 8, pp 229–245
40. Lanzillo NA, Breneman CM (2016) Predicting single-molecule conductance through machine learning. J Appl Phys 120:13
41. Wang L, Deng Y, Knight JL, Wu Y, Kim B, Sherman W, Shelley JC, Lin T, Abel R (2013) Modeling local structural rearrangements using FEP/REST: application to relative binding affinity predictions of CDK2 inhibitors. J Chem Theory Comput 9(2):1282–1293
42. Tadesse S, Anshabo AT, Portman N, Lim E, Tilley W, Caldon CE, Wang S (2020) Targeting CDK2 in cancer: challenges and opportunities for therapy. Drug Discovery Today 25(2):406–413
43. Green W, Burns J, Zalte AS, Abreu C, Sieg J, Feldmann C, Mathea M (2026) Deep Learning Foundation Models from Classical Molecular Descriptors
44. Fooladi H, Vu TNL, Mathea M, Kirchmair J (2025) Evaluating machine learning models for molecular property prediction: performance and robustness on out-of-distribution data. J Chem Inf Model 65(19):9871–9891
45. Kretschmer F, Seipp J, Ludwig M, Klau GW, Böcker S (2025) Coverage bias in small molecule machine learning. Nat Commun 16(1):554
46. Gapsys V, Pérez-Benito L, Aldeghi M, Seeliger D, Van Vlijmen H, Tresadern G, De Groot BL (2020) Large scale relative protein ligand binding affinities using non-equilibrium alchemy. Chem Sci 11(4):1140–1152
47. Baumann HM, Gapsys V, de Groot BL, Mobley DL (2021) Challenges encountered applying equilibrium and nonequilibrium binding free energy calculations. J Phys Chem B 125(17):4241–4261
48. Chang Y, Xu S, Ding K (2019) Tyrosine kinase 2 (TYK2) allosteric inhibitors to treat autoimmune diseases. ACS Publications 62:8951–8952
49. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749

50. Wu D, Kofke DA (2005) Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation. J Chem Phys *123* (5)
51. Cave-Ayland C, Skylaris C-K, Essex JW (2015) Direct validation of the single step classical to quantum free energy perturbation. J Phys Chem B 119(3):1017–1025
52. Liu S, Wu Y, Lin T, Abel R, Redmann JP, Summa CM, Jaber VR, Lim NM, Mobley DL (2013) Lead optimization mapper: automating free energy calculations for lead optimization. J Comput Aided Mol Des 27(9):755–770
53. Schoenmaker L, Jiskoot DA, Scheen J, Cheng E, Gapsys V, Hahn DF, Ries B, van Westen GJ, Mobley DL, Jespers W (2025) IMERGE-FEP: Improving Relative Free Energy Calculation Convergence with Chemical Intermediates. J Phys Chem B 129(9):2370–2379
54. Furui K, Shimizu T, Akiyama Y, Kimura SR, Terada Y, Ohue M (2025) PairMap: An Intermediate Insertion Approach for Improving the Accuracy of Relative Free Energy Perturbation Calculations for Distant Compound Transformations. J Chem Inf Model 65(2):705–721
55. Baumann HM, Dybeck E, McClendon CL, Pickard IV, Gapsys FC, Pérez-Benito V, Hahn L, Tresadern DF, Mathiowetz G, Mobley AM (2023) Broadening the scope of binding free energy calculations using a separated topologies approach. J Chem Theory Comput 19(15):5058–5076
56. Tsai H-C, Xu J, Guo Z, Yi Y, Tian C, Que X, Giese T, Lee T-S, York DM, Ganguly A (2024) Improvements in precision of relative binding free energy calculations afforded by the alchemical enhanced sampling (ACES) approach. J Chem Inf Model 64(18):7046–7055
57. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M (2019) Analyzing learned molecular representations for property prediction. J Chem Inf Model 59(8):3370–3388
58. Liu R, Wang H, Glover KP, Feasel MG, Wallqvist A (2018) Dissecting machine-learning prediction of molecular activity: is an applicability domain needed for quantitative structure–activity relationship models based on deep neural networks? J Chem Inf Model 59(1):117–126
59. Li Y, Liu R, Liu J, Luo H, Wu C, Li Z (2022) An open source graph-based weighted cycle closure method for relative binding free energy calculations. J Chem Inf Model 63(2):561–570
60. Cui D, Zhang BW, Tan Z, Levy RM (2019) Ligand binding thermodynamic cycles: hysteresis, the locally weighted histogram analysis method, and the overlapping states matrix. J Chem Theory Comput 16(1):67–79
61. Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW (2021) Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci 7(8):1356–1367
62. Scalia G, Grambow CA, Pernici B, Li Y-P, Green WH (2020) Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model 60(6):2697–2717
63. Heid E, McGill CJ, Vermeire FH, Green WH (2023) Characterizing uncertainty in machine learning for chemistry. J Chem Inf Model 63(13):4012–4029

64. Sørensen J, Bannan CC, Calabrò G, Jain V, Ovanesyan G, Smith A, Zhang S, Bayly CI, Darden TA, Geballe MT (2024) Orion® A Cloud-Native Molecular Design Platform. Comput Drug Discovery: Methods Appl 2:579–615

Figures

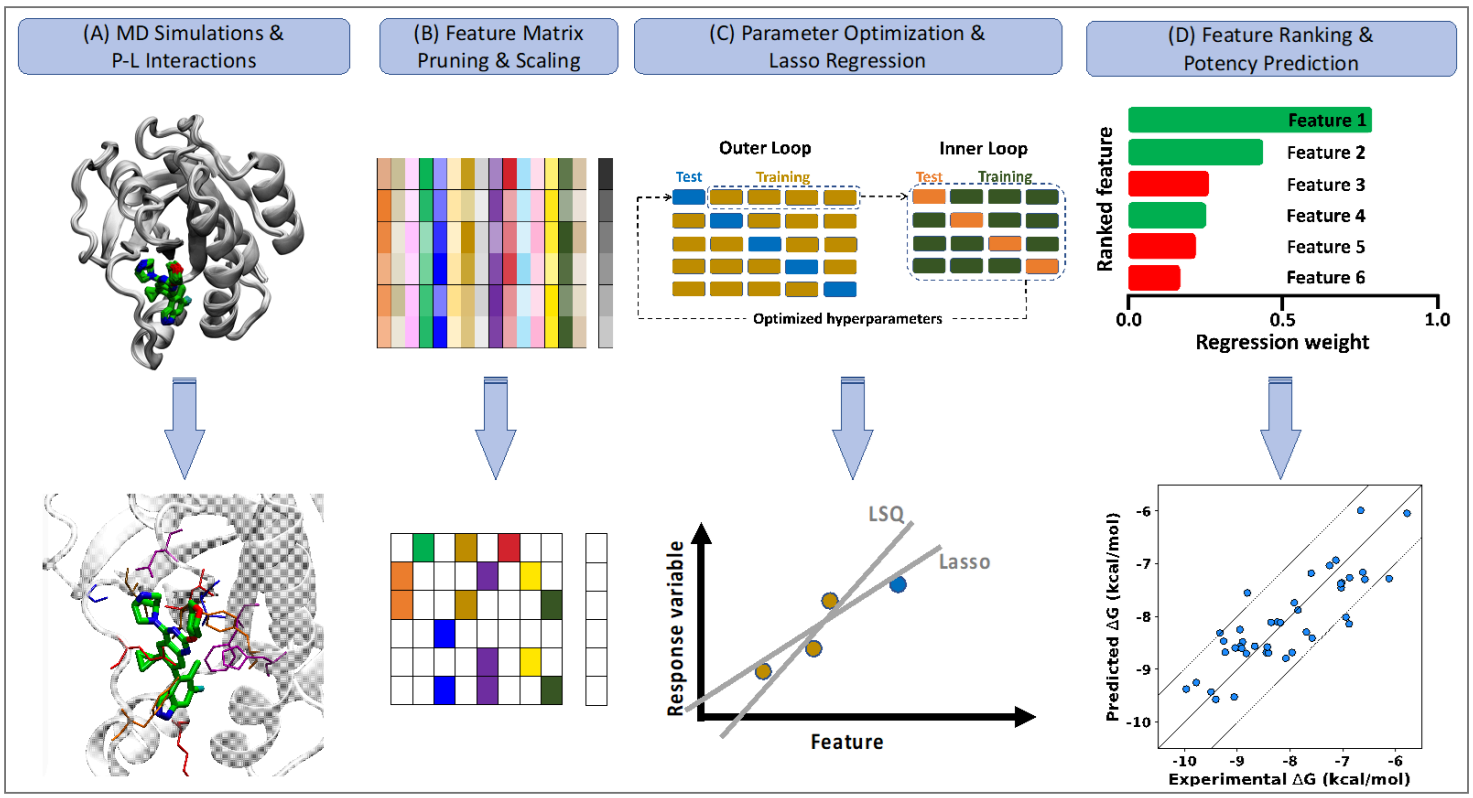


Figure 1

MDFitML takes a library of ligands with reasonable starting poses in a protein pocket, runs MD, and generates collated interaction fingerprints called SimFP. These interactions are then used as features to describe the ligand set and train a Lasso model against program relevant assay end-points such as potency. The Lasso model identifies top N SimFP interactions that can explain these potency differences. A simple linear regression model is then built using these top N features for final prediction. By mapping potency differences back into such transient pocket interactions observed in MD this approach offers unprecedent interpretability to prediction outcomes that are not always apparent from single docked poses of the ligand set.

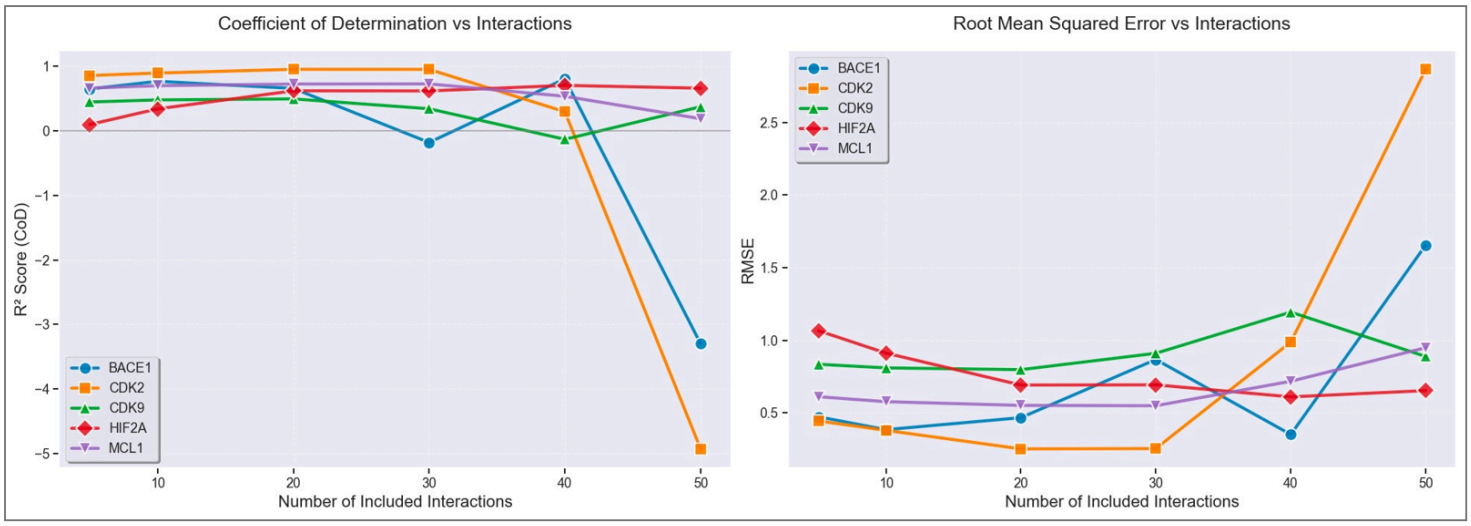


Figure 2

Impact of feature set size on model performance across public protein targets. Linear regression model performance evaluated using varying numbers of top-ranked Lasso-selected interactions for five protein targets from public datasets (BACE1, CDK2, CDK9, Hif-2α, and MCL-1). (Left) CoD scores and (Right) RMSE values.

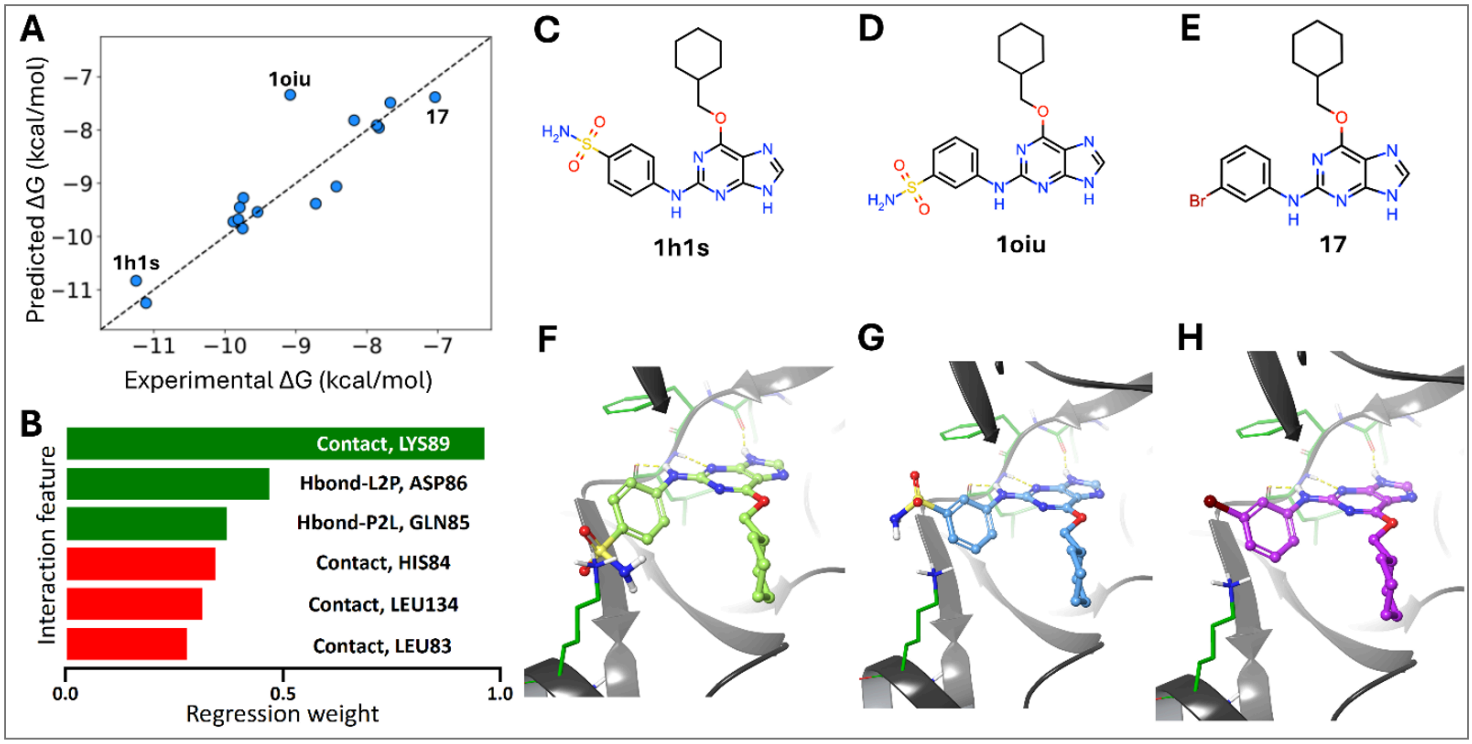


Figure 3

Explaining changes in binding affinities to CDK2 using SimFP. (A) The parity plot of binding free energies from MDFitML versus dG from measured Ki. The three distinct ligands shown in panels C-H are highlighted here. (B) The top six SimFP that were assigned the highest regression coefficients based on

the Lasso model are shown along with their contribution to strong (green) and weak (red) binding. (C, D, E) 2D-structures of three example ligands that bind strongly (1h1s), moderately (1oiu), and weakly (17) among the 16 ligands in this series. (F, G, H) 3D-visualization of the simulation starting structures for the three ligands bound to CDK2. In each panel, the ligand is shown in a ball-and-stick model, while the protein is visualized in cartoon with some binding-site side chains (dark green) in a stick model. Hbonds are illustrated as dashed. Note how contacts with LYS89 increase with binding affinity, while the Hbond donor-acceptor motif with the kinase hinge region is present for all three example ligands.

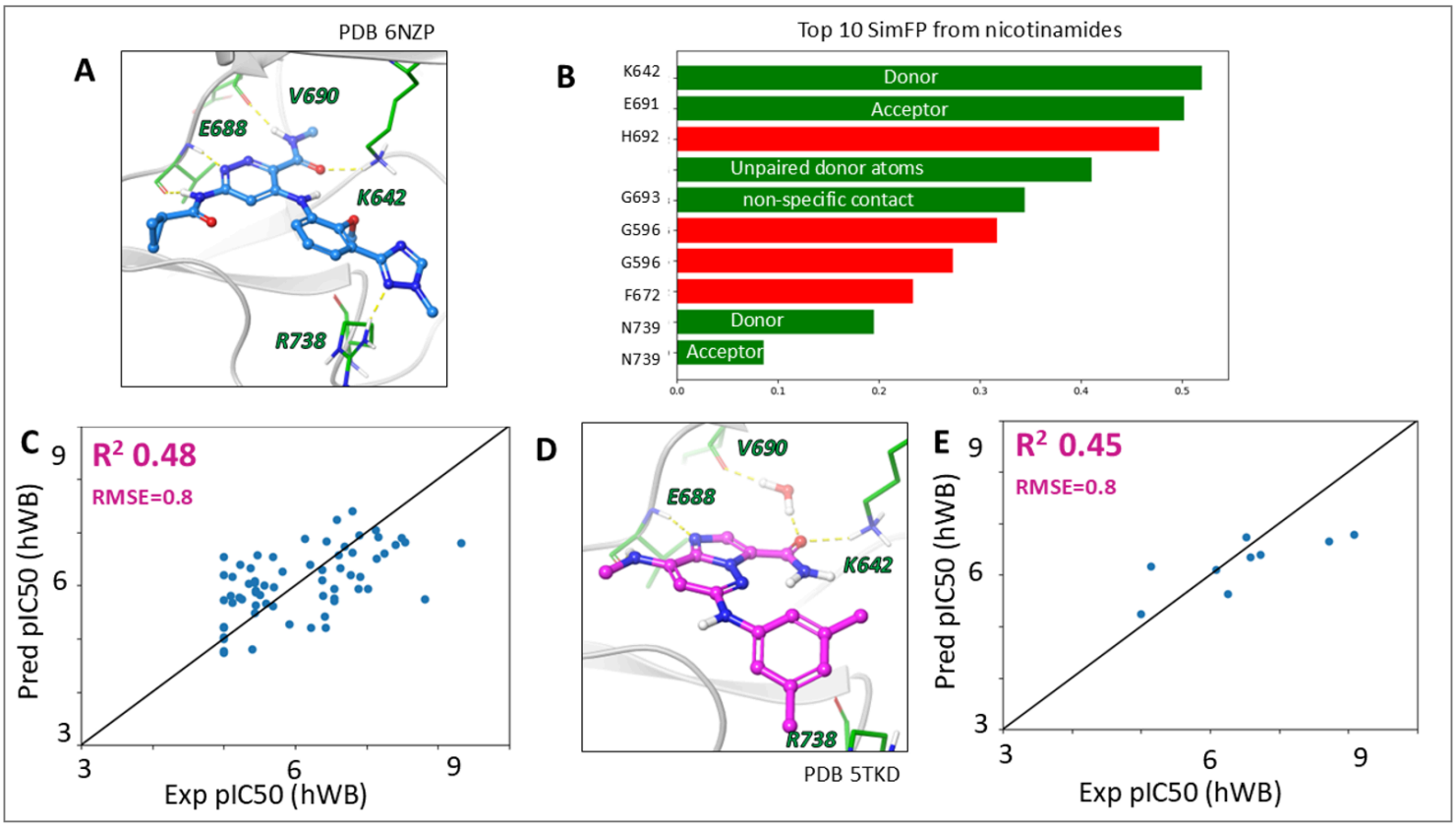


Figure 4

MDFitML regression model can be readily transferable to explain potency changes across different chemotypes targeting Tyk2. (A) Co-crystal structure of BMS-986165 bound to JH2 pocket of Tyk2 (PDB : 6NZP). (B) Top ten SimFP identified from MD simulations of 66 nicotinamide compounds based on the Lasso analysis and assigned the highest regression coefficients shown along with their contribution to strong (green) and weak (red) binding. (C) Parity plot of human whole blood (hWB) pIC50 from MDFitML regression model vs. measured. (D) Co-crystal structure of imidazopyrdizine compound 18 bound to same JH2 pocket as BMS-986165. (E) Parity plot shows decent correlation between predicted pIC50 using the same nicotinamide MDFitML regression model but this time predicting IZPs using their SimFP.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CDK2dataset.csv
- BACE1dataset.csv
- HiF2adataset.csv
- MCL1dataset.csv
- CDK9dataset.csv
- SIMDFitMLformatted.docx