
Look mom, no experimental data! Learning to score protein-ligand interactions from simulations

Michael Brocidiaco¹ James Wellnitz Konstantin I. Popov Alexander Tropsha
Eshelman School of Pharmacy
University of North Carolina at Chapel Hill
¹mixarcid@unc.edu

Abstract

Despite recent advances in protein-ligand structure prediction, deep learning methods remain limited in their ability to accurately predict binding affinities, particularly for novel protein targets dissimilar from the training set. In contrast, physics-based binding free energy calculations offer high accuracy across chemical space but are computationally prohibitive for large-scale screening. We propose a hybrid approach that approximates the accuracy of physics-based methods by training target-specific neural networks on molecular dynamics simulations of the protein in complex with random small molecules. Our method uses force matching to learn an implicit free energy landscape of ligand binding for each target. Evaluated on six proteins, our approach achieves competitive virtual screening performance using 100-500 μ s of MD simulations per target. Notably, this approach achieves state-of-the-art early enrichment when using the true pose for active compounds. These results highlight the potential of physics-informed learning for virtual screening on novel targets. We publicly release the code for this paper at <https://github.com/molecularmodelinglab/lfm> under the MIT license.

1 Introduction

Finding small molecule ligands that bind to a protein target is a critical step in early-stage drug discovery. However, this objective can be extremely challenging and expensive – the typical approach is to run a high-throughput screen (HTS), which can take months to set up and cost millions of dollars [16]. Identifying true binders faster and more cheaply is a major goal of computational drug discovery.

Traditionally, the challenge of predicting protein-ligand binding affinities is separated into two tasks: docking and scoring. Docking algorithms propose 3D binding poses of a ligand within a protein’s binding site, while scoring functions estimate the binding affinity of a given pose. While much recent progress has been made in the area of docking [13, 4, 56], scoring has lagged behind [58, 53, 10]. This is the primary reason for the unreliability of structure-based virtual screening (SBVS), wherein ligands from a large library are ranked according to their predicted binding affinity, and the top-scoring hits are chosen for experimental characterization. More recently, generative modeling has emerged as a potential alternative to SBVS; however, generative models are still typically evaluated and/or fine-tuned according to inaccurate scoring functions such as AutoDock Vina [51, 41, 39].

The difficulty of scoring arises from both data limitations and fundamental trade-offs in existing approaches. ML-based scoring functions often suffer from overfitting due to the scarcity of high-quality binding affinity data [58, 53]. On the other hand, physics-based methods exhibit a speed-accuracy trade-off: rigorous absolute binding free energy (ABFE) simulations provide accurate affinity estimates but require extensive computational resources, often taking many hours per

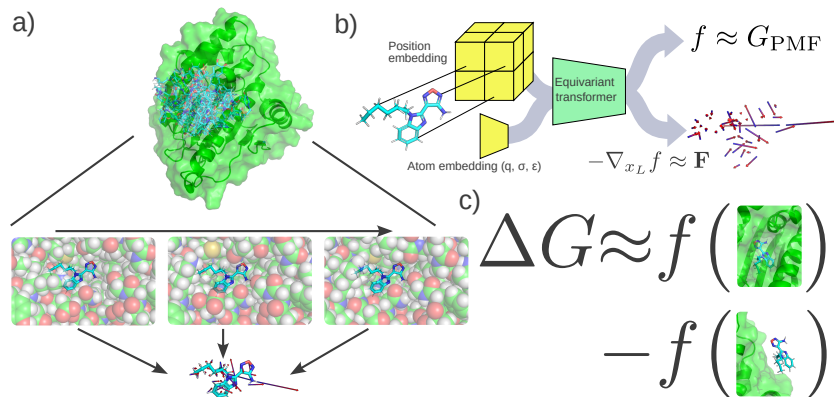


Figure 1: Method overview. a) For a particular protein target, we start by placing random small molecules in the binding site. For each ligand, we run an MD simulation of the protein-ligand complex, keeping the ligand atoms frozen relative to the binding site. We record the positions and intermolecular forces of the ligand atoms. b) Using this dataset, we train a neural network to approximate this binding PMF of the protein-ligand complex using force matching. c) In a virtual screen, we use this neural network to approximate the free energy of protein-ligand binding by subtracting the PMF of the ligand outside the binding site from the PMF of the docked ligand.

molecule even on GPUs [7, 24, 5]. In contrast, traditional fast scoring functions, such as those used by Vina and Glide [23], offer rapid predictions but are often unreliable, failing to achieve robust enrichment in virtual screening tasks [10, 32, 44].

Developing scoring methods that generalize beyond the available training data is especially important for the next generation of drugs. Recent advances in drug discovery have expanded interest beyond the traditional druggable proteome, targeting intrinsically disordered proteins [43], protein-protein interactions [38], and RNA tertiary structures [12] as promising avenues for next-generation therapeutics. The ability to efficiently identify small-molecule binders for these novel targets could greatly accelerate the development of new treatments. These novel target classes represent a formidable yet rewarding challenge for screening methods, as no current methodology can reliably find binders for such targets.

In this work, we propose combining the advantages of physics-based and machine learning models for protein-ligand scoring. Instead of trying to train a general binding affinity predictor to work on all protein targets, we seek to train a new model for each protein target of interest. We create a custom training set for each protein by running molecular dynamics (MD) simulations of random small molecules inside the binding site. We then use force matching to train a neural network on the free energy landscape of ligand binding. We emphasize that this approach requires no experimental data for the target in question except for its 3D structure (which can be estimated using structure-prediction methods such as AlphaFold [30]). We name this workflow ligand force matching (LFM).

2 Method

2.1 Overview

Invoking the Born-Oppenheimer approximation [9], a receptor-ligand system can be specified by the positions of the receptor atoms: x_R , the ligand atoms x_L , and the solvent atoms x_S . Throughout this paper, we use a coordinate system defined by the rigid-body alignment of the binding site α carbons to a reference structure. At equilibrium, these positions follow the Boltzmann distribution:

$$p(x_R, x_L, x_S) = \frac{1}{Z} e^{-\beta U(x_R, x_L, x_S)} \quad (1)$$

Here U is the potential energy of the system and $\beta = \frac{1}{k_B T}$ is the Boltzmann factor. In practice, we can approximate U with a fast and easily-evaluated sum of nonbonded and bonded energy terms.

When scoring protein-ligand complexes, we desire to predict the *binding free energy* $\Delta G_{\text{binding}} = G_{\text{bound}} - G_{\text{unbound}}$. The free energy G for a particular state is defined:

$$G_{\Omega} = -\frac{1}{\beta} \log \int_{\Omega} e^{-\beta U(x_L, x_R, x_S)} dx_L dx_R dx_S \quad (2)$$

Here Ω is the set of either all bound or all unbound poses, depending on the free energy in question.

The integral in this equation is computationally intractable, as protein-ligand systems can easily exceed 10K atoms. In practice, we can approximate $\Delta G_{\text{binding}}$ by sampling from many intermediate systems wherein the energy terms corresponding to the protein-ligand interactions are slowly attenuated; the final free energy is the work done during this transformation. This is the approach taken by ABFE protocols such as double decoupling [26] and the alchemical transfer method [57]. As mentioned above, such algorithms are accurate but slow.

As an alternative approach, we can condition G on the ligand coordinates x_L . This results in the *binding potential of mean force (PMF)*.

$$G_{\text{PMF}}(x_L) = -\frac{1}{\beta} \log \int_{-\infty}^{\infty} e^{-\beta U(x_L, x_R, x_S)} dx_R dx_S \quad (3)$$

As its name implies, the force this potential generates on the ligand atoms is equal to the ensemble average of the ligand forces from the full potential U .

$$-\nabla_{x_L} G_{\text{PMF}}(x_L) = -\left\langle \nabla_{x_L} U(x_L, x_R, x_S) \right\rangle_{x_R, x_S} \quad (4)$$

Estimating the full $\Delta G_{\text{binding}}$ is much easier if we have access to G_{PMF} , as it already integrates over the majority of the degrees of freedom. We can additionally quickly approximate the full free energy difference as the difference between G_{PMF} of a single bound and unbound ligand pose.

While G_{PMF} is still intractable to compute directly, we can easily estimate the forces $\mathbf{F} = -\nabla_{x_L} G_{\text{PMF}}$ via a single MD simulation.

In standard molecular mechanics force fields, U can be split into the ligand-ligand and ligand-environment components $U(x_L, x_R, x_S) = U_{LL}(x_L) + U_{LE}(x_L, x_R, x_S)$. $U_{LL}(x_L)$ is easy to calculate, so, in practice, we want to estimate the intermolecular PMF $\hat{G}_{\text{PMF}} = G_{\text{PMF}} - U_{LL}$, utilizing the intermolecular forces $\hat{\mathbf{F}}$.

We desire a model that can quickly approximate \hat{G}_{PMF} ; this should be able to generalize to new ligands but is specific to the current target in question. We do this by generating a dataset $\mathcal{D} = (x_L, a_L, \mathbf{F})$ of random small molecules in random positions across the binding site. Here, a_L are nonbonded parameters $[\sigma, \epsilon, q]$ needed to specify the Lennard-Jones and Coulombic energies for the force field.

Finally, we train a neural network f to approximate \hat{G}_{PMF} (up to a constant) by minimizing the force-matching loss [54]:

$$\mathcal{L}_F = \left(-\nabla_{x_L} f(x_L, a_L) - \hat{\mathbf{F}} \right)^2 \quad (5)$$

This loss function has previously been used to train coarse-grained and implicit solvent force fields [15, 31].

When using the trained model in practice, we only want to evaluate the model on “reasonable” poses given to it by the docking program (that are presumably relatively low-energy), as well as poses outside the binding site (to subtract off the PMF in solvent). Therefore, when generating data, we only need to sample such reasonable poses, while ensuring that we sample along a path between the center of the binding site and somewhere outside. We describe one such method below.

2.2 Data generation

We begin by running a 100 ns MD simulation of the target receptor to produce an ensemble of conformations. We then compute the “exit point” x_E , defined as a point close to the pocket center that is at least c (here, 9 Å) away from any receptor atom. We assume that a molecule centered at this exit point is far enough away from the receptor that its interactions are negligible. We then generate each datapoint as follows:

We sample a random ligand from the ZINC database [48]. We select a pose for this ligand by first generating a conformer with RDKit [1]. We then uniformly sample a point x between the pocket center and the exit point. We then place the center of mass of the ligand according to a normal distribution centered at x and with standard deviation σ_P (here 6 Å). The ligand is rotated randomly around this position.

We then select a random apo protein conformation (with solvent) from its MD trajectory. The ligand is placed in the pocket according to the position specified above. The ligand now likely has steric clashes with the protein and solvent. Additionally, the ligand is likely in an unreasonable pose that will not help training. To rectify these two problems, we run a short (100 ps) MD simulation of the complex, slowly adding the interaction terms between the ligand and its environment according to a soft-core alchemical potential [8] defined by the sterics interaction term λ_s and the electrostatics term λ_c . We first drive λ_s from 0 to 1 over 90 ps and then drive λ_c from 0 to 1 over the remaining 10 ps. Additionally, we ensure that the center of mass of the ligand remains close to its starting position via a harmonic restraint ($k = 10 \text{ kJ mol}^{-1} \text{ nm}^{-2}$). Thus, by the end of this alchemical simulation, all steric clashes should be removed. Additionally, the ligand is now in a more energetically favorable conformation while its center of mass is still in the relatively same position.

Now that we have a ligand position that meets our criteria, we run the final simulation for 4 ns to determine the mean force on the ligand atoms. During this simulation, we freeze the ligand atoms. Additionally, to ensure that the ligand remains in the same place relative to the protein binding site, we restrain the protein atoms using a harmonic force between the pocket atoms and the reference pocket atoms, projected so that it only affects the target’s rigid-body degrees of freedom (see Appendix A.4).

During the simulation, we log the intramolecular forces acting on the ligand atoms every 10 ps. We compute the final mean force after the equilibration time (here, 2.5 ns).

All data in this paper was generated via OpenMM [21] with the Amber ff14SB [34] and Espaloma [55] force fields.

2.3 Model Architecture and Training

To predict the PMF, our model needs to have embeddings for both the ligand atom features and the position of the atoms within the binding site. Note that this task is not invariant with respect to the ligand coordinates, so a purely equivariant architecture will not work.

To create the ligand atom embedding h_L , we utilize an MLP over the atom parameters a . The ligand coordinate embedding is done via cubic spline interpolation over uniformly spaced 3D grids. We define several such grids with various spacings and hidden dimensions. We use `torch-cubic-spline-grids` [3] to define an embedding for each atom over each grid. The final ligand coordinate embedding is an MLP over the concatenated grid embeddings for each atom.

The final combined embedding h_c is an MLP over the concatenated ligand embeddings, coordinate embeddings, and the multiplication of the two embeddings.

These embeddings are then fed into an equivariant transformer from TorchMDNet [50]. The final energy is the sum of individual atom energies.

We train the model with force matching. In addition to just the force-matching loss, we also include the mean-squared-error loss for the force on the ligand center of mass $\mathcal{L}_{\bar{F}}$ and torque of the ligand center of mass $\mathcal{L}_{\bar{T}}$. This is done because we care more about the relative PMFs of the ligand in different positions. The final loss is thus $\mathcal{L} = \lambda_F \mathcal{L}_F + \lambda_{\bar{F}} \mathcal{L}_{\bar{F}} + \lambda_{\bar{T}} \mathcal{L}_{\bar{T}}$. In this paper, we use $\lambda_F = 1.0$, $\lambda_{\bar{F}} = 50.0$, and $\lambda_{\bar{T}} = 5.0$.

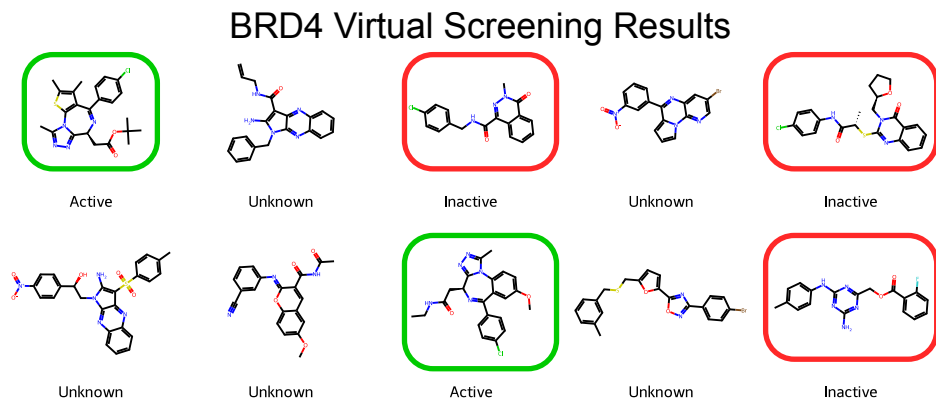


Figure 2: Results from the in-house virtual screen with BRD4. We could not find 5 of the compounds in the collection, so they were not tested; of the three that were tested, none were found to bind.

2.4 Model use

To score a docked ligand, we first compute the total energy (intramolecular energy + model energy $f(x)$) for each docked pose. We select the minimum energy pose and compute the final score as $\Delta G_{\text{binding}} \approx f(x) - f(x_{\text{solv}})$. Here, x_{solv} is the ligand outside the binding site. (To compute this, we re-center the ligand at the exit point defined during data generation.)

This approximation of binding affinity assumes a rigid ligand and therefore neglects the ligand reorganization energy of binding. A more accurate (albeit slower) approximation would be to explicitly compute the free energy between the solvated and bound structures, e.g., by using the alchemical transfer method [57] over the PMF. We leave this to future work.

3 Results

3.1 BRD4 virtual screening

We first tried this method on the well-studied cancer target BRD4, utilizing an older variant of the dataset generation (outlined in Appendix A.3) to generate 50K datapoints. After achieving promising results on retrospective data (shown below), we used our model to screen a 170K in-house library and selected the top-predicted binders for experimental validation.

As seen in Figure 2, two of the top ten compounds selected are known BRD4 binders. Encouraged, we elected to test the remaining 8 for BRD4 binding via Isothermal Calorimetry (ITC). Unfortunately, we were only able to acquire 3 of the compounds from the library; of these, none demonstrated measurable binding affinity. Thus, though we failed to discover novel binders, we did rediscover two known BRD4 binders.

3.2 Benchmarking on all targets

We additionally tried this method on 5 other targets (MCL1, ESR1, MDM2, CDK2, and HSP90). We chose these targets for a combination of factors, namely their small size (enabling faster and cheaper MD simulations) and their large and diverse number of known binders, both with and without co-crystal structures (enabling us to robustly test virtual screening performance).

For each protein, we compiled a set of protein-ligand cocrystal structures from the PLINDER [20] dataset. We chose an arbitrary structure to initialize our method from this set, along with a set of aligned ligands with 3D poses. We also searched ChEMBL [25] for known binders and selected a diverse subset of small molecules with known activity but no known binding pose. Finally, we generated up to 50 property-matched decoys for each active molecule, following the example of DUD-E [37]. Our primary evaluation metric is the ability of LFM to distinguish known binders from these decoys, ideally with high early enrichment.

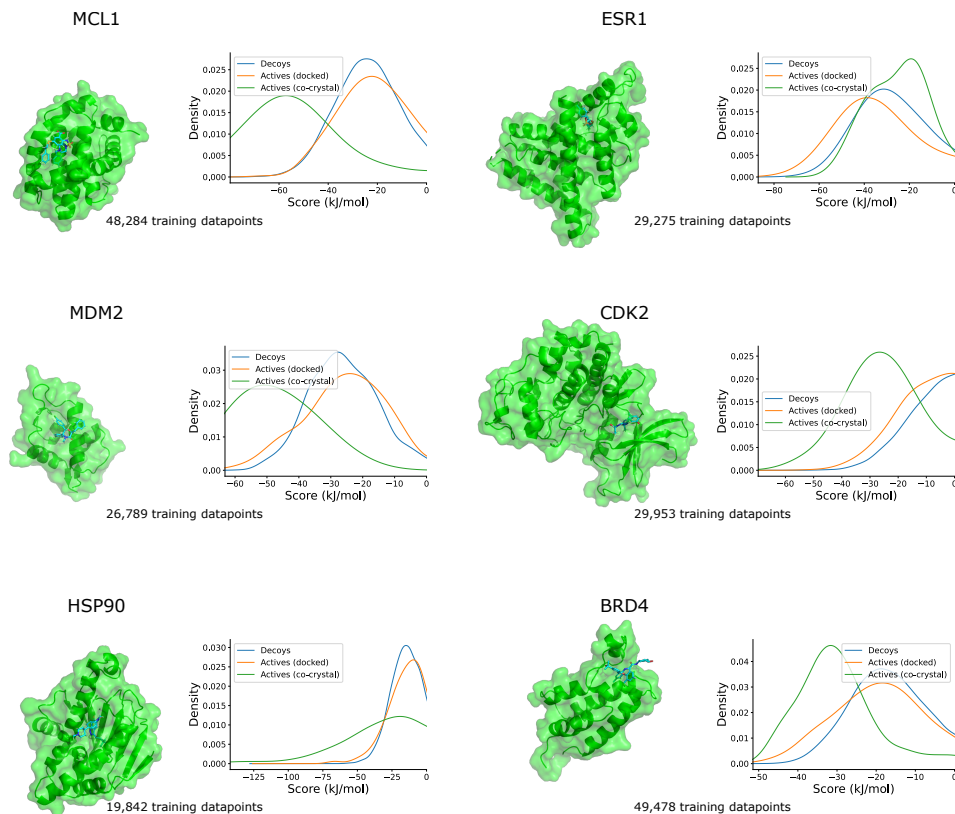


Figure 3: Each benchmarking target used in this study, co-crystallized with a representative ligand. For each target, we show the distribution of LFM scores for actives and decoys docked with DIFFDOCK, as well as the co-crystal structures of the actives with known structure. In general, the scores of the docked actives are lower than the docked decoys, showcasing the value of the LFM models for virtual screening. We see a more dramatic score shift for the co-crystallized active molecules, showcasing the pose-dependence of the LFM models.

We ran the LFM workflow for each protein, ensuring that no molecule used for training the model was within 0.3 Tanimoto similarity coefficient to any known binder from ChEMBL (2048-bit, radius 4 Morgan fingerprint [36]). This enables us to robustly test the ability of our method to generalize across chemical space. We generated 25K-60K datapoints per target, utilizing the cloud resources of VastAI [2]. We split the dataset into 80:10:10 train:validation:test sets. This amounts to 100-500 μ s of training MD simulation time per protein, for a cost of around \$1K each. We trained each model for up to 300 epochs on an NVIDIA L40 GPU, saving the model that achieved the best validation mean force loss. We used the AdamW [33] optimizer with a learning rate of 10^{-4} .

None of these targets is especially novel, meaning that traditional screening methods are likely to work adequately well on them already. We would ideally like to test our method on targets that are out-of-domain for current methods. However, by their nature, novel targets lack large numbers of diverse binders that can robustly allow us to test the performance of this method retrospectively. Thus, we leave the evaluation of LFM on novel target classes to future prospective studies.

We used both Uni-Dock [59] and DIFFDOCK-L [14, 13] to dock the binders and decoys for each target. Uni-Dock implements a faster GPU-enabled version of the algorithm used by Vina. DIFFDOCK has demonstrated improved docking accuracy relative to traditional methods such as Vina. It should perform especially well on these well-studied targets since it was trained on all of them. The use of these two docking methods will demonstrate how improved docking performance impacts the performance of scoring models.

Table 1: Median metrics for each model when using known co-crystal poses for the active compounds and using Uni-Dock or DIFFDOCK to dock the decoy compounds. We additionally report docking accuracy at a 2 Å cutoff when using the model to rerank the poses proposed by Uni-Dock or DIFFDOCK.

Docking method	Model	EF_{\max}^B	$EF_{1\%}^B$	AUC	% < 2 Å
Uni-Dock	Vina	3.7 [1.0, 13]	0.0 [0.0, 0.0]	0.36 ± 0.09	13 ± 5.0 %
	GNINA	120 [16, 230]	12 [5.9, 23]	0.79 ± 0.05	13 ± 5.4 %
	LFM	290 [120, 620]	28 [18, 44]	0.88 ± 0.07	19 ± 5.2 %
DIFFDOCK	Vina	25 [12, 72]	8.5 [4.3, 22]	0.69 ± 0.06	63 ± 6.8 %
	GNINA	120 [28, 230]	20 [11, 28]	0.83 ± 0.06	56 ± 6.9 %
	LFM	640 [68, 1000]	37 [24, 49]	0.86 ± 0.06	64 ± 8.2 %

We then used the LFM models, along with the baseline scoring models Vina [51] and GNINA [35] to score and rerank the resulting poses. GNINA is a 3D CNN-based model that has previously demonstrated state-of-the-art performance in virtual screening [49]. Since its neural networks were trained using 3-fold cross-validation with clustered splits [22], we used an ensemble of neural networks that were not trained on the cluster containing the target in question.

The primary metric for evaluating the scoring models is the Bayes enrichment factor EF^B [10]. Given a cutoff fraction χ and a set of model scores S on both active and decoy compounds, this metric is an estimate of the probability ratio $\frac{P(\text{Active}|S>S_\chi)}{P(\text{Active})}$. This is the probability that the top-scoring hits from a virtual screen will bind, relative to selecting compounds at random. We are primarily interested in the maximum achievable enrichment EF_{\max}^B , though we also report the enrichment at $\chi = 1\%$. It is worth noting that, since these metrics depend so heavily on the tail of the score distribution, they are highly sensitive to noise in the data. Thus, we should pay close attention to the (asymmetric) error bars (computed via bootstrapping).

3.3 Benchmarking results

As shown in Table 1, the LFM models perform better along all metrics when utilizing true crystal poses for the actives molecules and either Uni-Dock or DIFFDOCK for the decoy molecules. Additionally, the LFM models can more accurately pick out “correct” docked structures (within 2 ÅRMSD) from a set of proposed structures.

The story becomes more complicated when we use docked poses for both the actives and decoys. As we see in Table 2, the LFM models perform about as well in terms of enrichment as the other models when using both Uni-Dock or DIFFDOCK for both the actives and decoys. Additionally, GNINA is the best at classifying actives from inactives in terms of AUC.

We suggest that these results can be explained by the increased pose-sensitivity of the LFM models. The score given by GNINA (here, the CNN_VS output) is the product of an affinity score and a pose score; thus, it is more likely to rank an active above a decoy even if both their poses are deemed incorrect. This especially explains the higher AUCs achieved by GNINA relative to its enrichment factors, as these incorrectly docked actives will not be in the top selected hits despite contributing to AUC.

On the other hand, the LFM models are, by design, extremely sensitive to the input pose. Indeed, these models achieve state-of-the-art pose selection performance when reranking poses from both docking methods tested. We also see a more dramatic increase in enrichment when using more accurate docked poses (DIFFDOCK versus Uni-Dock), and even state-of-the-art performance when utilizing crystal poses.

We additionally hypothesize that DIFFDOCK gives unrealistic poses for the decoy molecules. This is evidenced in Table 1, as all models show improved performance discriminating actives with crystal poses from decoys with DIFFDOCK poses than decoys with Uni-Dock poses. Clearly, the models are giving lower scores to the decoys docked by DIFFDOCK. This makes sense when we consider that DIFFDOCK was only trained on binders from the PDB.

Table 2: Median metrics for each model when using Uni-Dock or DIFFDOCK for both active and decoy poses.

Docking method	Model	EF_{\max}^B	$EF_{1\%}^B$	AUC
Uni-Dock	Vina	9.7 [3.4, 15]	1.9 [0.86, 2.7]	0.54 ± 0.03
	GNINA	6.9 [4.2, 14]	3.7 [2.1, 4.5]	0.59 ± 0.02
	LFM	4.6 [3.3, 14]	2.0 [1.2, 3.1]	0.52 ± 0.02
DIFFDOCK	Vina	16 [4.7, 25]	3.5 [2.0, 4.4]	0.60 ± 0.02
	GNINA	13 [7.7, 27]	5.8 [3.9, 7.3]	0.68 ± 0.02
	LFM	14 [7.8, 33]	6.2 [3.6, 8.7]	0.58 ± 0.03
DIFFDOCK (actives)	Vina	1.0 [1.0, 2.6]	0.35 [0.0, 0.62]	0.23 ± 0.02
Uni-Dock (decoys)	GNINA	7.9 [4.9, 17]	3.9 [2.3, 5.0]	0.65 ± 0.02
	LFM	13 [7.9, 31]	5.8 [2.8, 8.3]	0.55 ± 0.04

With this in mind, we hypothesized that using DIFFDOCK to dock the actives and Uni-Dock to dock the decoys is a better test case for how scoring models perform in the regime where docking models are all around better. Indeed, we see that LFM models are able to perform best in this case. This is consistent with our hypothesis that these models are more pose-sensitive than the others.

Though prospective use of LFM will, of course, use docked rather than crystal poses, we suggest that this pose sensitivity of the LFM models corresponds to the fact that they have learned more generalizable features. Actual binding affinities are highly dependent on the specific intermolecular interactions being made by the protein-ligand complex. Thus, while the LFM models perform as well as the baselines on these relatively “in-domain” targets, we hypothesize that LFM will be more reliable when used for more novel targets.

LFM inference took an average of 2.5 s on an L40 GPU per molecule, which is fast enough to use in large-scale screening.

Overall, we see that our LFM models demonstrate remarkable early enrichment for three of the six targets. The other three, however, proved more difficult for the method.

3.4 TBXT virtual screening

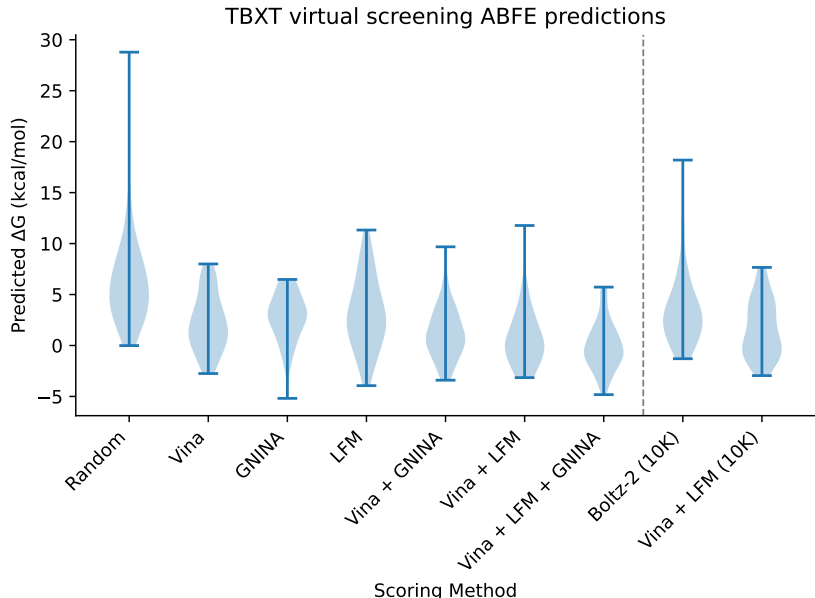
To test how well LFM and the baselines work on novel binding sites in a real-world setting, we ran a virtual screening campaign for the TBXT transcription factor. This is an important and currently undrugged target implicated in chordoma and other cancers [28, 45]. We ran a 4 μ s MD simulation of TBXT and identified a potential cryptic binding site. Since this is a novel binding site on an already novel target, it represents an incredibly difficult but realistic test case for virtual screening methods. While a couple (weak) binders for this target exist in the PDB, none bind to the site we identified.

We used Uni-Dock to generate binding poses for 1M random compounds from the Enamine REAL set [27]. We used the baseline models as well as an LFM model trained on 50K datapoints to rescore these compounds and ran ABFE calculations on the top 50 compounds proposed by each method. We used the method proposed by Ding [17] to provide robust uncertainty estimates for each prediction; since these predictions had high variance, we used the upper bound 95% confidence interval as a conservative estimate of the binding free energy. This can sometimes give unrealistically high ΔG estimates (in general, the true binding ΔG will be less than 0 kcal/mol), but these unrealistic estimates are more often given for poor binders with unstable binding poses anyway (as the current ABFE code assumes a relatively stable binding pose). As a control, we also ran these calculations on 50 random compounds. See Appendix A.6 for more details on the free energy calculations.

While this benchmark is far from perfect, as we are using predicted values rather than experimental binding affinities, we argue that such a benchmark still has value. As mentioned earlier, ABFE calculations are known to rank-order ligands well and are quite effective at discriminating actives from decoys [7, 24, 5].

We noticed that the top-scoring compounds proposed by the LFM method sometimes had strange poses on the side of the pocket. We reasoned that this could be an artifact of the training procedure,

Figure 4: Distribution of predicted binding affinities after running ABFE simulations on the top 50 results from each scoring method.



as such poses are unlikely to be representative of ligands in the training set; perhaps a predicted low-energy region existed there by accident. We hypothesized that combining LFM and Vina models would produce better top-scoring compounds by grounding the model and preventing these sort of “hallucinations”. Thus we also tested the top 50 compounds when we combined the ranks of Vina and LFM scores (and only keeping compounds whose top-scoring Vina and LFM conformer was within 2 Å). We also tested this same procedure with GNINA and Vina, using the same reasoning. Finally, we tested a combination of all three models.

We also used the recently released Boltz-2 as a baseline [40]. Because Boltz-2 takes much longer to run than the other models (around 5 minutes rather than a couple of seconds), we were unable to run it on the entire library of 1M compounds. Instead, we used it to screen 10K compounds. We also ran the Vina and LFM method on these same compounds as a direct comparison. We did not benchmark more models on this 10K library due to the expense of running ABFE calculations; our primary goal was to test if the accuracy of Boltz-2 was enough to compensate for its speed. It’s worth noting that we did not benchmark Boltz-2 against the previous well-characterized targets since it was (likely) trained on all of them.

The results of this screening benchmark are shown in Figure 4 and Table 3. All models are significantly better than random guessing (as measured by mean predicted ΔG). Curiously, Vina, GNINA, and the LFM model all perform similarly, and ensembling multiple models helps a lot. The combination of all three models performs the best. Though the triple combination is not a statistically significant improvement over the Vina + LFM combination ($p = 0.097$), it is over the Vina + GNINA combination ($p = 0.008$).

Notably, the Vina + LFM combination is a significant improvement over Boltz-2 when using a 10K library ($p = 0.003$), despite running approximately 100x faster.

4 Discussion

We presented ligand force matching, a new approach to protein-ligand scoring that utilizes molecular dynamics simulations to train a per-target neural network. We demonstrated the potential of this method on 6 protein targets, utilizing 100-500 μs of MD simulation data per target. The LFM models achieve a state-of-the-art early enrichment when using the true co-crystal poses for the active compounds, and the performance remains competitive when using docked poses.

Table 3: TBXT virtual screening results. For each model, we report the mean and median predicted binding affinities (as given by our ABFE calculations). We also return the percentage below 0 kcal/mol.

Library size	Model	Median pred. $\Delta G \downarrow$	Mean pred. $\Delta G \downarrow$	% < 0 kcal/mol \uparrow
1M	Vina	1.8 ± 1.0	2.2 ± 0.77	22 ± 12
	GNINA	2.8 ± 0.68	2.7 ± 0.65	12 ± 9.0
	LFM	2.6 ± 1.2	3.1 ± 0.99	22 ± 11
	Vina + GNINA	0.89 ± 0.70	1.2 ± 0.73	34 ± 13
	Vina + LFM	0.22 ± 0.75	0.68 ± 0.81	44 ± 14
	Vina + LFM + GNINA	-0.15 ± 0.65	-0.0027 ± 0.71	56 ± 13
10K	Boltz-2	2.8 ± 0.89	3.4 ± 0.98	12 ± 9.0
	Vina + LFM	0.79 ± 1.2	1.6 ± 0.77	36 ± 13

Ligand force-matching is a general method for developing virtual screening models for novel targets. The specific workflow used in this paper is just one possible variant of the method, and there are many possible improvements that could further boost model performance. For instance, there are many possible neural network architectures to try beyond the equivariant transformer used in this work. Of particular interest is the Gaussian process approach used by Vandermause et al. [52], which could enable robust uncertainty estimation for our predictions.

On the data generation side, there are many possible things to try. For instance, the current workflow presented only uses a single protonation state for the protein when generating data, but, in reality, histidines in the binding site often change protonation state based on the bound ligand. A more careful approach would be to assign protonation state separately for each pose in the training set before MD simulation, or to do more advanced MD simulations at constant pH [18, 11]. Additionally, there could be slow timescale movements that the 4 ns simulations could not capture. Longer simulations during data generation (or enhanced sampling methods) could alleviate this. Finally, there are inherent inaccuracies in the molecular mechanics (MM) force field. New neural network potentials (NNPs) have recently shown great success at quickly approximating quantum mechanical energies, and even in improving protein-ligand binding affinity predictions [46, 19, 42]. Current NNPs are still much slower than MM force fields, but we could use an NNP to reweigh the forces from an MM simulation.

The most important limitation of this method is the requirement for large-scale molecular dynamics simulations for each target studied. While this is generally worth the cost (high-throughput screening still costs orders of magnitude more money), future work should be done in creating MD datasets across the proteome to train a general-purpose network.

While we focused on the application of this method to scoring, a similar method could also be used for docking, as a ligand’s pose in the binding site should be the minimum of the free energy surface. The LFM models presented already outperform the baselines in terms of reranking poses from Uni-Dock and DIFFDOCK. Going forward, we intend to train smaller versions of the network (similar to the simple energy functions used by Vina or Glide) that would allow for quickly minimizing the predicted free energy. This would allow for predicting poses that require greater receptor rearrangement, a task at which current docking methods struggle.

We intend to use this method for prospective hit discovery for novel targets, and we hope that other researchers do the same. There is also much work to be done in testing the limits of this method. How much binding site flexibility can it handle? How well does it work for more difficult target classes, such as intrinsically disordered proteins, protein-protein interactions, or even RNA targets?

In conclusion, we believe that ligand force matching has the potential to transform early-stage drug discovery for novel targets. We hope that future researchers continue to test and refine this method so it can achieve its potential.

We release the code for the dataset generation, model training, and model benchmarking at <https://github.com/molecularmodelinglab/lfm> under the MIT license.

Acknowledgments and Disclosure of Funding

The authors thank Dr. David Koes, Dr. Yinglong Miao, Enes Kelestemur, Rishabh Dey, Kushal Koirala, and Nyssa Tucker for insightful comments and discussion.

References

- [1] RDKit: Open-source cheminformatics, . URL <https://www.rdkit.org/>.
- [2] Rent GPUs | Vast.ai, . URL <https://vast.ai/>.
- [3] teamtomo/torch-cubic-spline-grids, May 2025. URL <https://github.com/teamtomo/torch-cubic-spline-grids>. original-date: 2023-01-19T11:57:39Z.
- [4] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Židek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL <https://www.nature.com/articles/s41586-024-07487-w>. Publisher: Nature Publishing Group.
- [5] Matteo Aldeghi, Alexander Heifetz, Michael J. Bodkin, Stefan Knapp, and Philip C. Biggin. Accurate calculation of the absolute free energy of binding for drug molecules. *Chemical Science*, 7(1):207–218, December 2015. ISSN 2041-6539. doi: 10.1039/C5SC02678D. URL <https://pubs.rsc.org/en/content/articlelanding/2016/sc/c5sc02678d>. Publisher: The Royal Society of Chemistry.
- [6] Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics*, 31(13):2214–2216, July 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv082. URL <https://doi.org/10.1093/bioinformatics/btv082>.
- [7] Irfan Alibay, Aniket Magarkar, Daniel Seeliger, and Philip Charles Biggin. Evaluating the use of absolute binding free energy in the fragment optimisation process. *Communications Chemistry*, 5(1):1–13, September 2022. ISSN 2399-3669. doi: 10.1038/s42004-022-00721-4. URL <https://www.nature.com/articles/s42004-022-00721-4>. Publisher: Nature Publishing Group.
- [8] Thomas C. Beutler, Alan E. Mark, René C. van Schaik, Paul R. Gerber, and Wilfred F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical Physics Letters*, 222(6):529–539, June 1994. ISSN 0009-2614. doi: 10.1016/0009-2614(94)00397-1. URL <https://www.sciencedirect.com/science/article/pii/0009261494003971>.
- [9] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484, 1927. ISSN 1521-3889. doi: 10.1002/andp.19273892002. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19273892002>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.19273892002>.
- [10] Michael Brocidiacono, Konstantin I. Popov, and Alexander Tropsha. An Improved Metric and Benchmark for Assessing the Performance of Virtual Screening Models, March 2024. URL <http://arxiv.org/abs/2403.10478>. arXiv:2403.10478 [q-bio].

- [11] Yunjie Chen and Benoît Roux. Constant-pH Hybrid Nonequilibrium Molecular Dynamics–Monte Carlo Simulation Method. *Journal of Chemical Theory and Computation*, 11(8):3919–3931, August 2015. ISSN 1549-9618. doi: 10.1021/acs.jctc.5b00261. URL <https://doi.org/10.1021/acs.jctc.5b00261>. Publisher: American Chemical Society.
- [12] Jessica L. Childs-Disney, Xueyi Yang, Quentin M. R. Gibaut, Yuquan Tong, Robert T. Batey, and Matthew D. Disney. Targeting RNA structures with small molecules. *Nature Reviews Drug Discovery*, 21(10):736–762, October 2022. ISSN 1474-1784. doi: 10.1038/s41573-022-00521-4. URL <https://www.nature.com/articles/s41573-022-00521-4>. Publisher: Nature Publishing Group.
- [13] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking, October 2022. URL <http://arxiv.org/abs/2210.01776>. arXiv:2210.01776 [physics, q-bio].
- [14] Gabriele Corso, Arthur Deng, Benjamin Fry, Nicholas Polizzi, Regina Barzilay, and Tommi Jaakkola. Deep Confident Steps to New Pockets: Strategies for Docking Generalization, February 2024. URL <http://arxiv.org/abs/2402.18396>. arXiv:2402.18396 [q-bio].
- [15] Amaury Coste, Ema Slejko, Julija Zavadlav, and Matej Praprotnik. Developing an Implicit Solvation Machine Learning Model for Molecular Simulations of Ionic Media. *Journal of Chemical Theory and Computation*, 20(1):411–420, January 2024. ISSN 1549-9618. doi: 10.1021/acs.jctc.3c00984. URL <https://doi.org/10.1021/acs.jctc.3c00984>. Publisher: American Chemical Society.
- [16] John W Davies, Meir Glick, and Jeremy L Jenkins. Streamlining lead discovery by aligning *in silico* and high-throughput screening. *Current Opinion in Chemical Biology*, 10(4):343–351, August 2006. ISSN 1367-5931. doi: 10.1016/j.cbpa.2006.06.022. URL <https://www.sciencedirect.com/science/article/pii/S136759310600086X>.
- [17] Xinqiang Ding. Bayesian Multistate Bennett Acceptance Ratio Methods. *Journal of Chemical Theory and Computation*, 20(5):1878–1888, March 2024. ISSN 1549-9618. doi: 10.1021/acs.jctc.3c01212. URL <https://doi.org/10.1021/acs.jctc.3c01212>. Publisher: American Chemical Society.
- [18] Serena Donnini, Florian Tegeler, Gerrit Groenhof, and Helmut Grubmüller. Constant pH Molecular Dynamics in Explicit Solvent with λ -Dynamics. *Journal of Chemical Theory and Computation*, 7(6):1962–1978, June 2011. ISSN 1549-9618. doi: 10.1021/ct200061r. URL <https://doi.org/10.1021/ct200061r>. Publisher: American Chemical Society.
- [19] Timothy T. Duignan. The Potential of Neural Network Potentials. *ACS Physical Chemistry Au*, 4(3):232–241, May 2024. doi: 10.1021/acsphyschemau.4c00004. URL <https://doi.org/10.1021/acsphyschemau.4c00004>. Publisher: American Chemical Society.
- [20] Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, Emanuele Rossi, Guoqing Zhou, Srimukh Veccham, Clemens Isert, Yuxing Peng, Prabindh Sundareson, Mehmet Akdel, Gabriele Corso, Hannes Stärk, Gerardo Tauriello, Zachary Carpenter, Michael Bronstein, Emine Kucukbenli, Torsten Schwede, and Luca Naef. PLINDER: The protein-ligand interactions dataset and evaluation resource, July 2024. URL <https://www.biorxiv.org/content/10.1101/2024.07.17.603955v3>. Pages: 2024.07.17.603955 Section: New Results.
- [21] Peter Eastman, Raimondas Galvelis, Raúl P. Peláez, Charles R. A. Abreu, Stephen E. Farr, Emilio Gallicchio, Anton Gorenko, Michael M. Henry, Frank Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A. Mitchell, Vijay S. Pande, João PGLM Rodrigues, Jaime Rodriguez-Guerra, Andrew C. Simmonett, Sukrit Singh, Jason Swails, Philip Turner, Yuanqing Wang, Ivy Zhang, John D. Chodera, Gianni De Fabritiis, and Thomas E. Markland. OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The Journal of Physical Chemistry B*, 128(1):109–116, January 2024. ISSN 1520-6106. doi: 10.1021/acs.jpcb.3c06662. URL <https://doi.org/10.1021/acs.jpcb.3c06662>. Publisher: American Chemical Society.

- [22] Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *Journal of Chemical Information and Modeling*, 60(9):4200–4215, September 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00411. URL <https://doi.org/10.1021/acs.jcim.0c00411>. Publisher: American Chemical Society.
- [23] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. *Journal of Medicinal Chemistry*, 47(7):1739–1749, March 2004. ISSN 0022-2623. doi: 10.1021/jm0306430. URL <https://doi.org/10.1021/jm0306430>. Publisher: American Chemical Society.
- [24] Haohao Fu, Yan Zhou, Xiang Jing, Xueguang Shao, and Wensheng Cai. Meta-Analysis Reveals That Absolute Binding Free-Energy Calculations Approach Chemical Accuracy. *Journal of Medicinal Chemistry*, 65(19):12970–12978, October 2022. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.2c00796. URL <https://doi.org/10.1021/acs.jmedchem.2c00796>. Publisher: American Chemical Society.
- [25] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(Database issue):D1100–D1107, January 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr777. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245175/>.
- [26] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal*, 72(3):1047–1069, March 1997. ISSN 0006-3495. doi: 10.1016/S0006-3495(97)78756-3. URL [https://www.cell.com/biophysj/abstract/S0006-3495\(97\)78756-3](https://www.cell.com/biophysj/abstract/S0006-3495(97)78756-3). Publisher: Elsevier.
- [27] Oleksandr O. Grygorenko, Dmytro S. Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E. Gubina, and Yurii S. Moroz. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience*, 23(11):101681, November 2020. ISSN 2589-0042. doi: 10.1016/j.isci.2020.101681. URL <https://www.sciencedirect.com/science/article/pii/S2589004220308737>.
- [28] Duane H. Hamilton, Justin M. David, Charli Dominguez, and Claudia Palena. Development of cancer vaccines targeting brachyury, a transcription factor associated with tumor EMT. *Cells, tissues, organs*, 203(2):128–138, 2017. ISSN 1422-6405. doi: 10.1159/000446495. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC5381518/>.
- [29] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), June 2023. URL <http://arxiv.org/abs/1606.08415>. arXiv:1606.08415 [cs].
- [30] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL <https://www.nature.com/articles/s41586-021-03819-2>. Number: 7873 Publisher: Nature Publishing Group.
- [31] Paul Katzberger and Sereina Riniker. A general graph neural network based implicit solvation model for organic molecules in water. *Chemical Science*, 15(28):10794–10802, July 2024. ISSN 2041-6539. doi: 10.1039/D4SC02432J. URL <https://pubs.rsc.org/en/content/articlelanding/2024/sc/d4sc02432j>. Publisher: The Royal Society of Chemistry.

- [32] Peter Kolb and John J. Irwin. Docking screens: right for the right reasons? *Current Topics in Medicinal Chemistry*, 9(9):755–770, 2009. ISSN 1568-0266. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383315/>.
- [33] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv:1711.05101 [cs, math]*, January 2019. URL <http://arxiv.org/abs/1711.05101>. arXiv: 1711.05101.
- [34] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, August 2015. ISSN 1549-9618. doi: 10.1021/acs.jctc.5b00255. URL <https://doi.org/10.1021/acs.jctc.5b00255>. Publisher: American Chemical Society.
- [35] Andrew T. McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. GNINA 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 13(1):43, June 2021. ISSN 1758-2946. doi: 10.1186/s13321-021-00522-2. URL <https://doi.org/10.1186/s13321-021-00522-2>.
- [36] H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965. ISSN 0021-9576. doi: 10.1021/c160017a018. URL <https://doi.org/10.1021/c160017a018>. Publisher: American Chemical Society.
- [37] Michael M. Mysinger, Michael Carchia, John J. Irwin, and Brian K. Shoichet. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, July 2012. ISSN 0022-2623. doi: 10.1021/jm300687e. URL <https://doi.org/10.1021/jm300687e>. Publisher: American Chemical Society.
- [38] Hossam Nada, Yongseok Choi, Sungdo Kim, Kwon Su Jeong, Nicholas A. Meanwell, and Kyeong Lee. New insights into protein–protein interaction modulators in drug discovery and therapeutic advance. *Signal Transduction and Targeted Therapy*, 9(1):1–32, December 2024. ISSN 2059-3635. doi: 10.1038/s41392-024-02036-3. URL <https://www.nature.com/articles/s41392-024-02036-3>. Publisher: Nature Publishing Group.
- [39] Joseph A. Newman, Alice Douangamath, Setayesh Yadzani, Yuliana Yosaatmadja, Antony Aimon, José Brandão-Neto, Louise Dunnett, Tyler Gorrie-stone, Rachael Skyner, Daren Fearon, Matthieu Schapira, Frank von Delft, and Opher Gileadi. Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase. *Nature Communications*, 12(1):4848, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25166-6. URL <https://www.nature.com/articles/s41467-021-25166-6>.
- [40] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction, June 2025. URL <https://www.biorxiv.org/content/10.1101/2025.06.14.659707v1>. Pages: 2025.06.14.659707 Section: New Results.
- [41] Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets, May 2022. URL <http://arxiv.org/abs/2205.07249>. arXiv:2205.07249 [cs].
- [42] Francesc Sabanés Zariquiey, Raimondas Galvelis, Emilio Gallicchio, John D. Chodera, Thomas E. Markland, and Gianni De Fabritiis. Enhancing Protein–Ligand Binding Affinity Predictions Using Neural Network Potentials. *Journal of Chemical Information and Modeling*, 64(5):1481–1485, March 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c02031. URL <https://doi.org/10.1021/acs.jcim.3c02031>. Publisher: American Chemical Society.
- [43] Suman Saurabh, Karthik Nadendla, Shubh Sanket Purohit, Ponnurengam Malliappan Sivakumar, and Sibel Cetinel. Fuzzy Drug Targets: Disordered Proteins in the Drug-Discovery Realm. *ACS Omega*, 8(11):9729–9747, March 2023. doi: 10.1021/acsomega.2c07708. URL <https://doi.org/10.1021/acsomega.2c07708>. Publisher: American Chemical Society.

- [44] Thomas Scior, Andreas Bender, Gary Tresadern, José L. Medina-Franco, Karina Martínez-Mayorga, Thierry Langer, Karina Cuanalo-Contreras, and Dimitris K. Agrafiotis. Recognizing Pitfalls in Virtual Screening: A Critical Review. *Journal of Chemical Information and Modeling*, 52(4):867–881, April 2012. ISSN 1549-9596. doi: 10.1021/ci200528d. URL <https://doi.org/10.1021/ci200528d>. Publisher: American Chemical Society.
- [45] Tanaz Sharifnia, Mathias J. Wawer, Ting Chen, Qing-Yuan Huang, Barbara A. Weir, Ann Sizemore, Matthew A. Lawlor, Amy Goodale, Glenn S. Cowley, Francisca Vazquez, Christopher J. Ott, Joshua M. Francis, Slim Sassi, Patricia Cogswell, Hadley E. Sheppard, Tinghu Zhang, Nathanael S. Gray, Paul A. Clarke, Julian Blagg, Paul Workman, Josh Sommer, Francis Hornicek, David E. Root, William C. Hahn, James E. Bradner, Kwok K. Wong, Paul A. Clemons, Charles Y. Lin, Joanne D. Kotz, and Stuart L. Schreiber. Small-molecule targeting of brachyury transcription factor addiction in chordoma. *Nature medicine*, 25(2):292–300, February 2019. ISSN 1078-8956. doi: 10.1038/s41591-018-0312-3. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC6633917/>.
- [46] J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science*, 8(4):3192–3203, March 2017. ISSN 2041-6539. doi: 10.1039/C6SC05720A. URL <https://pubs.rsc.org/en/content/articlelanding/2017/sc/c6sc05720a>. Publisher: The Royal Society of Chemistry.
- [47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. ISSN 1533-7928. URL <http://jmlr.org/papers/v15/srivastava14a.html>.
- [48] Teague Sterling and John J. Irwin. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, November 2015. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00559. URL <https://doi.org/10.1021/acs.jcim.5b00559>. Publisher: American Chemical Society.
- [49] Jocelyn Sunseri and David Ryan Koes. Virtual Screening with Gnina 1.0. *Molecules*, 26(23):7369, January 2021. ISSN 1420-3049. doi: 10.3390/molecules26237369. URL <https://www.mdpi.com/1420-3049/26/23/7369>. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute.
- [50] Philipp Thölke and Gianni De Fabritiis. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials, April 2022. URL <http://arxiv.org/abs/2202.02541>. arXiv:2202.02541 [cs].
- [51] Oleg Trott and Arthur J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of computational chemistry*, 31(2):455–461, January 2010. ISSN 0192-8651. doi: 10.1002/jcc.21334. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041641/>.
- [52] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1):1–11, March 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-0283-z. URL <https://www.nature.com/articles/s41524-020-0283-z>. Publisher: Nature Publishing Group.
- [53] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *Journal of Medicinal Chemistry*, 65(11):7946–7958, June 2022. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.2c00487. URL <https://doi.org/10.1021/acs.jmedchem.2c00487>. Publisher: American Chemical Society.
- [54] Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charron, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Central Science*, 5(5):755–767, May 2019. ISSN 2374-7943. doi:

- 10.1021/acscentsci.8b00913. URL <https://doi.org/10.1021/acscentsci.8b00913>. Publisher: American Chemical Society.
- [55] Yuanqing Wang, Josh Fass, Benjamin Kaminow, John E. Herr, Dominic Rufa, Ivy Zhang, Iván Pulido, Mike Henry, and John D. Chodera. End-to-End Differentiable Molecular Mechanics Force Field Construction. *Chemical Science*, 13(41):12016–12033, 2022. ISSN 2041-6520, 2041-6539. doi: 10.1039/D2SC02739A. URL <http://arxiv.org/abs/2010.01196>. arXiv:2010.01196 [physics].
- [56] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 Democratizing Biomolecular Interaction Modeling, November 2024. URL <https://www.biorxiv.org/content/10.1101/2024.11.19.624167v1>. Pages: 2024.11.19.624167 Section: New Results.
- [57] Joe Z. Wu, Solmaz Azimi, Sheenam Khuttan, Nanjie Deng, and Emilio Gallicchio. Alchemical Transfer Approach to Absolute Binding Free Energy Estimation. *Journal of Chemical Theory and Computation*, 17(6):3309–3319, June 2021. ISSN 1549-9618. doi: 10.1021/acs.jctc.1c00266. URL <https://doi.org/10.1021/acs.jctc.1c00266>. Publisher: American Chemical Society.
- [58] Jincai Yang, Cheng Shen, and Niu Huang. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Frontiers in Pharmacology*, 11, 2020. ISSN 1663-9812. URL <https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2020.00069>.
- [59] Yuejiang Yu, Chun Cai, Jiayue Wang, Zonghua Bo, Zhengdan Zhu, and Hang Zheng. Uni-Dock: GPU-Accelerated Docking Enables Ultralarge Virtual Screening. *Journal of Chemical Theory and Computation*, 19(11):3336–3345, June 2023. ISSN 1549-9618. doi: 10.1021/acs.jctc.2c01145. URL <https://doi.org/10.1021/acs.jctc.2c01145>. Publisher: American Chemical Society.

A Technical Appendices and Supplementary Material

A.1 Dataset generation

For BRD4, the first target that the LFM workflow was tested on, we generated the ligand poses for the dataset slightly differently. Instead of randomly initializing the ligand in the binding site, we used QuickVina [6] to initialize the pose in a random protein conformation from the apo trajectory. We then generated a random translation that took the ligand outside the binding site, and a random rotation. We then produced the final translation as a uniformly random interpolation between the docked and random poses. Finally, we froze the ligand in this pose and alchemically added it to the random apo protein frame. The intuition was to sample the full space of ligand poses while sampling more around “good” poses given by QuickVina, similar to how datapoints are generated in flow-matching models. While this strategy worked well for BRD4, it quickly became apparent that it would not work well in general. The binding sites for the other proteins in the benchmarking set tended to close during the apo trajectory, meaning that QuickVina was rarely able to find “reasonable” ligand poses. Hence we elected to go for the less biased data generation approach presented above.

For CDK2 and HSP90, we used a slightly different ligand restraint algorithm when generating the data. Specifically, we restrained the distance between the ligand center of mass and the protein center of mass to be the same as the one when the ligand was randomly positioned. We changed it to a full restraint on the ligand center of mass to the original ligand center of mass to keep the ligand closer to the “exit path,” thus (hopefully) being more data efficient.

A.2 Model architecture

Each model was a 3-layer Equivariant Transformer architecture as implemented in TorchMDNet [50] with a hidden dimension of 128. To produce input atom embeddings for the model, we first

generated a “receptor embedding” from 3 interpolated cubice spline grids with hidden dimension [64, 256, 512] and spacing [2 Å 4 Å 8 Å]. The grid was sized so that it included all the residues in the protein binding site and the “exit point” for the binding site. We concatenated these interpolated embedding and applied a Dropout layer [47] before producing a 128-dimensional hidden vector. We additionally generated a “ligand embedding” as a linear layer mapping the atom nonbonded parameters (partial charge q and Lennard-Jones ϵ and σ) to a 128-dimensional hidden vector before applying Dropout. Finally, we produced a “both embedding” by multiplying both the embeddings. The final atom embedding was a linear layer and Dropout over the concatenated ligand, receptor, and both embeddings to produce a 128-dimensional hidden vector. We used a SiLU activation function for all the layers described [29], and a dropout rate of 0.2 for all the Dropout layers.

A.3 BRD4 virtual screen

To run the BRD4 virtual screen, we used Uni-Dock to dock the entire in-house library before rescoring the ligands with the LFM model (unlike the results presented in this paper, we minimized the docked poses with the LFM model before rescoring). We nominated the top ten virtual hits with less than 6 rotatable bonds, ensuring that no hit with within 0.3 Tanimoto similarity to any other hit.

A.4 Rigid-body restraining force

When generating the dataset to train the LFM models, we ensure that the receptor remains in the reference frame defined by the pocket α carbons by the following force:

$$F_i = \overline{F} + \frac{\overline{T}_i \times x_i}{|x|^2}$$

Here \overline{F} and \overline{T} are the mean force and torque over the pocket α carbons for the unprojected harmonic force $\hat{F}_i = -2k(x_i - \hat{x}_i)$, where \hat{x} are the original pocket α carbon positions. In this work, we used $k = 10^4 \text{ kJ mol}^{-1} \text{ nm}^{-2}$.

A.5 Model scoring

Before using the LFM models to score docked poses, we minimize the intramolecular energy of the ligand (with harmonic restraints to the original docked coordinates with $k=5 \text{ kJ mol}^{-1} \text{ nm}^{-2}$). We do this because the docked poses may have bond lengths and angles that are out-of-domain relative to the poses that the model has been trained on.

A.6 Free energy calculations

We used the double-decoupling ABFE method [26] with nonequilibrium switching to estimate absolute binding free energies. We used a vacuum intermediate ligand state that was harmonically restrained to its binding pose (taken after a 5 ns simulation) with $k=10 \text{ kJ mol}^{-1} \text{ nm}^{-2}$). We equilibrated the ligand-only, apo, and holo simulations for 10 ns before sampling frames every 500 ps for the switching. We used 30 2 ns forward and backward switches when decoupling the ligand from the receptor, and 60 2 ns forward and backward switches when decoupling the ligand from solvent.