ToxBench: A Binding Affinity Prediction Benchmark with AB-FEP-Calculated Labels for Human Estrogen Receptor Alpha

Meng Liu [1 *] **Karl Leswing** [2 *] **Simon K. S. Chu** [1] **Farhad Ramezanghorbani** [1] **Griffin Young** [2] **Gabriel Marques** [2] **Prerna Das** [2] **Anjali Panikar** [2] **Esther Jamir** [2] **Mohammed Sulaiman Shamsudeen** [2] **K. Shawn Watts** [2] **Ananya Sen** [2] **Hari Priya Devannagari** [2] **Edward B. Miller** [2] **Muyun Lihan** [2] **Howook Hwang** [2] **Janet Paulsen** [1] **Xin Yu** [1] **Kyle Gion** [1] **Timur Rvachov** [1] **Emine Kucukbenli** [1] **Saee Gopal Paliwal** [1]

Abstract

Protein-ligand binding affinity prediction is essential for drug discovery and toxicity assessment. While machine learning (ML) promises fast and accurate predictions, its progress is constrained by the availability of reliable data. In contrast, physics-based methods such as absolute binding free energy perturbation (AB-FEP) deliver high accuracy but are computationally prohibitive for high-throughput applications. To bridge this gap, we introduce ToxBench, the first large-scale AB-FEP dataset designed for ML development and focused on a single pharmaceutically critical target, Human Estrogen Receptor Alpha (ERα). ToxBench contains 8,770 ERα-ligand complex structures with binding free energies computed via AB-FEP with a subset validated against experimental affinities at 1.75 *kcal/mol* RMSE, along with non-overlapping ligand splits to assess model generalizability. Using ToxBench, we further benchmark state-of-the-art ML methods, and notably, our proposed DualBind model, which employs a dual-loss framework to effectively learn the binding energy function. The benchmark results demonstrate the superior performance of DualBind and the potential of ML to approximate AB-FEP at a fraction of the computational cost.

1. Introduction

Fast and accurate prediction of protein-ligand binding affinity is fundamental for modern drug discovery. The ability to reliably estimate how tightly a small molecule binds to its target protein enables the prioritization of drug candidates and the early identification of off-target interactions that could lead to adverse effects (Meng et al., 2011; Pagadala et al., 2017; Chen et al., 2018; Yang et al., 2019b; Vamathevan et al., 2019; Beroza et al., 2022; Sadybekov & Katritch, 2023). Machine learning (ML) models offer the potential for rapid and accurate predictions, but their success critically depends on the availability of reliable datasets for training and evaluation. Unfortunately, obtaining large-scale and high-quality experimental affinity data remains a significant bottleneck due to the resource-intensive nature of experimental affinity measurements.

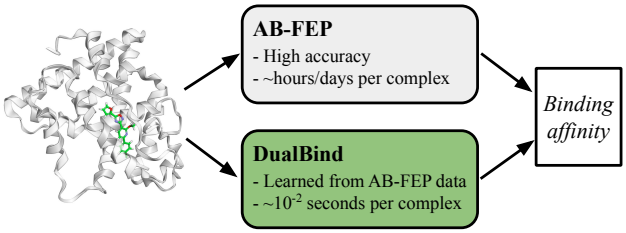

Figure 1. **An illustration of the ToxBench task.** ML models, such as DualBind, are trained on high-fidelity AB-FEP data to predict protein-ligand binding affinities several orders of magnitude faster than original AB-FEP calculations.

Existing affinity datasets have attempted to address this challenge but are shown to suffer from inherent biases that constrain model generalizability. Particularly, ML models are typically trained on the PDBBind dataset (Wang et al., 2004; Liu et al., 2017) and evaluated on the CASF-2016 benchmark (Su et al., 2018). However, recent studies indicate that models developed using these datasets tend to learn dataset-specific biases rather than truly capturing the underlying protein-ligand interactions (Volkov et al., 2022; Li et al., 2024; Durant et al., 2025). Notably, as detailed in Section 2, models trained solely on ligand or protein features, without

[*]Equal contribution [1]NVIDIA [2]Schrödinger. Correspondence to: Meng Liu <menliu@nvidia.com>, Karl Leswing <karl.leswing@schrodinger.com>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

modeling protein-ligand interactions, have achieved unexpectedly competitive performance on CASF-2016 (Yang et al., 2020; Volkov et al., 2022). This underscores the limitations of current benchmarks and highlights the urgent need for a more robust testbed that reflects real-world challenges.

In contrast to data-driven ML approaches, absolute protein-ligand binding free energy calculation through free energy perturbation (AB-FEP) has emerged as a highly accurate method. (Jorgensen et al., 1988; Aldeghi et al., 2016; Chen et al., 2023). By employing rigorous all-atom molecular dynamics (MD) simulations in explicit solvent, AB-FEP produces binding free energy calculations with accuracy comparable to experimental assays, which typically have uncertainties around 1.0 *kcal/mol* (Ross et al., 2023; Davis et al., 2011). Notably, the AB-FEP implementation in Schrödinger's FEP+ (Sch, 2021) achieves a root mean square error (RMSE) of $\sim$1.1 *kcal/mol* against experimental affinities in validation studies (Chen et al., 2023). However, due to extensive MD simulations, the AB-FEP calculation is computationally intensive and can often take hours to days to complete for a single complex system, thus limiting its practical use in high-throughput applications.

In response to the above challenges, in this paper, we introduce ToxBench, the first AB-FEP dataset centered on Human Estrogen Receptor Alpha (ERα), a pharmaceutically critical target. ERα is central to endocrine signaling and a key target for both therapeutic development and toxicity assessment, as its modulation is linked to various adverse outcomes, such as reproductive disorders and hormone-dependent cancers (Lee et al., 2013; La Merrill et al., 2020; Miziak et al., 2023). In particular, ToxBench comprises 8,770 ERα-ligand complexes with binding free energies computed via AB-FEP and partially validated against available experimental measurements. By incorporating non-overlapping ligand splits and concentrating on a single, pharmaceutically critical target, ToxBench closely aligns with real-world structure-based virtual screening scenarios, where extensive ligand libraries are assessed against one critical target. Therefore, ToxBench serves as a realistic testbed for developing and evaluating ML models for binding affinity prediction.

Using ToxBench, we conduct a comprehensive benchmarking study in which we train and evaluate representative ML models for binding affinity prediction. In addition, we propose a novel model, DualBind, which integrates supervised mean squared error (MSE) loss with unsupervised denoising score matching (DSM) loss to effectively learn the binding energy function. Our experimental results highlight the superior performance of DualBind and demonstrate that carefully designed ML models have the potential to approximate AB-FEP at a fraction of the computational cost, as illustrated in Figure 1.

Our contributions can be summarized as follows:

- **ToxBench Dataset**: We present ToxBench, the first large-scale AB-FEP dataset focused on the critical ERα target. ToxBench is designed as a realistic testbed for development of binding affinity prediction models.
- **DualBind Model**: We propose DualBind, a novel deep learning framework that synergistically combines MSE and DSM losses, leading to superior performance.
- **Benchmarking Study**: We conduct benchmarking experiments for state-of-the-art ML models on ToxBench, establishing robust baselines to guide future research in binding affinity prediction.

The dataset is available via `https://huggingface.co/datasets/karlleswing/toxbench` and the DualBind implementation can be found at `https://github.com/NVIDIA-Digital-Bio/dualbind`.

2. Related Work

Binding Affinity Prediction. Binding affinity prediction methods generally fall into three categories: scoring functions, physics-based approaches, and ML models. Conventional scoring functions (Böhm, 1994; Head et al., 1996; Eldridge et al., 1997; Böhm, 1998; Gohlke et al., 2000; Wang et al., 2002), usually based on empirical terms or knowledge-based analyses, offer rapid estimates but often lack accuracy, particularly when applied outside their training domains. In contrast, physics-based approaches, such as MM-PBSA (Kollman et al., 2000; Homeyer & Gohlke, 2012), MM-GBSA (Still et al., 1990; Gohlke et al., 2003; Gohlke & Case, 2004), and AB-FEP (Jorgensen et al., 1988; Gilson et al., 1997; Boresch et al., 2003), deliver more reliable predictions but are computationally extensive due to the high cost of molecular dynamics (MD) simulations and solvent modeling. In contrast, ML methods aim to achieve both speed and accuracy. Recent advances have introduced a variety of data-driven ML models for binding affinity prediction, built on sequence-based (Öztürk et al., 2018; Yuan et al., 2022), graph-based (Nguyen et al., 2021; Moon et al., 2022), or 3D structure-based (Jiménez et al., 2018; Jiang et al., 2021; Lu et al., 2022) representations and neural networks. Our proposed DualBind model, which considers spatial information, falls into the 3D structure-based category. Given the vast literature in this field, we refer readers to surveys such as Meng et al. (2011); Meli et al. (2022); Liu et al. (2024) for a comprehensive review of existing binding affinity prediction methods.

Existing Datasets and Limitations. The effectiveness of ML models for protein-ligand binding affinity prediction heavily depends on the quality of the datasets used for training and evaluation. Specifically, PDBBind (Wang et al.,

2004; Liu et al., 2017), which includes ∼20k protein-ligand complex structures with affinity measurements collected from various sources, and its 285-complex core set which is also known as CASF-2016 (Su et al., 2018), have become the de facto benchmarks in the field. ML models are typically trained on PDBBind and evaluated on the CASF-2016 benchmark. However, recent work by Volkov et al. (2022) demonstrates that, under this setup, models using only ligand or protein features, without modeling protein-ligand interactions, perform surprisingly well and modeling interactions does not provide clear advantage. In fact, Volkov et al. (2022)'s experiments reveal that a model trained solely on ligand graphs achieves a Pearson correlation coefficient (R_p) of 0.749 and an RMSE of 1.567, outperforming the model trained on protein-ligand interaction graphs (R_p = 0.687, RMSE = 1.605). This finding has been independently confirmed by multiple studies (Yang et al., 2020; Wang & Dokholyan, 2022; Durant et al., 2025). Together, these studies suggest that inherent biases in existing datasets cause models to rely on dataset-specific ligand-only or protein-only signals, rather than truly learning the underlying protein-ligand interactions. As a result, such trained models often fail to generalize to novel complexes (Volkov et al., 2022; Scantlebury et al., 2023).

Notably, Volkov et al. (2022) also defines the density of a training set as

$$D = \frac{\#\text{observed complexes}}{\#\text{proteins} \times \#\text{ligands}},$$

and reports $D < 0.05$ for PDBBind, meaning that over 95% of all possible protein–ligand pairs lack training affinity labels. Experimental studies from Volkov et al. (2022) indicate that, in such sparse regimes, models often exploit ligand-only or protein-only features as shortcuts, rather than learning protein-ligand interactions, and that increasing density is an attractive path to increase model generalizability. This insight motivates ToxBench. By focusing on a single target (ERα) and providing several thousands of AB-FEP-labeled complexes, ToxBench achieves full density for this receptor. This encourages ML models to truly learn protein-ligand interactions, thus improving their ability to predict affinities for novel ligands. We argue that such a task design aligns well with the practical target-based virtual screening scenario.

3. ToxBench Benchmark

Dataset Generation. The construction of the ToxBench dataset started with obtaining a set of ligand compounds targeting the Human Estrogen Receptor Alpha (ERα). The compound dataset was constructed by integrating data from the ChEMBL (Zdrazil et al., 2024) database and the Diverse Unbiased Validation-Extended (DUD-E) (Mysinger et al., 2012) dataset. An initial set of 473 compounds with reported assay values was obtained from ChEMBL. From this collection, 67 experimental binding affinity values were manually curated and standardized to units of *kcal/mol*. These specific values were prioritized due to their high data quality and consistent reporting as relative binding affinities. Furthermore, the binding affinity data for three specific ChEMBL compounds (CHEMBL234638, CHEMBL236086 and CHEMBL236718), initially reported in *Ki*, were validated and converted to *kcal/mol*.

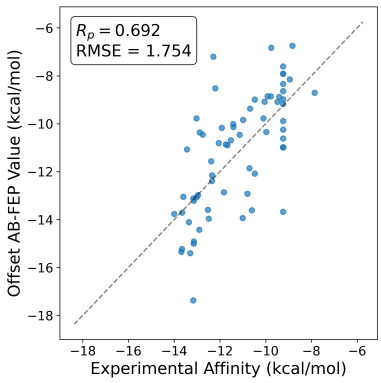

Figure 2. **AB-FEP calculations *vs.* experimental affinities**. This Compares AB-FEP calculated binding affinities with the corresponding 67 curated experimental binding affinities.

The structural representation of ERα in ToxBench was derived from three Protein Data Bank (PDB) (Berman et al., 2000) entries, which were refined using the Schrödinger Protein Preparation Wizard (Sastry et al., 2013). To account for the conformational diversity of ERα, PDB structure 1ERE was selected as representative of the agonist-bound conformation, while PDB structure 3ERT was selected to represent the antagonist-bound conformation. Additionally, PDB entry 1SJ0, which captures ERα in another distinct conformational state, was included to further diversify the protein templates.

Ligand preparation was conducted using Schrödinger LigPrep with Epik7 (Johnston et al., 2023; Schrödinger, LLC, 2024). For ChEMBL-derived ligands, a protocol involving Induced Fit Docking/Molecular Dynamics (IFD/MD) (Miller et al., 2021) was employed, followed by Absolute Binding Free Energy Perturbation (AB-FEP) (Chen et al., 2023) simulations performed for 1 ns. For ligands sourced from the DUD-E dataset, protein-ligand poses were generated using GLIDE-WS (Halgren et al., 2004; Murphy et al., 2016), also followed by 1 ns AB-FEP simulations.

Agreement of Calculations with Experimental Data. To assess the accuracy of our AB-FEP calculations, we compared them against the 67 curated experimental binding affinities. To obtain a single representative computational

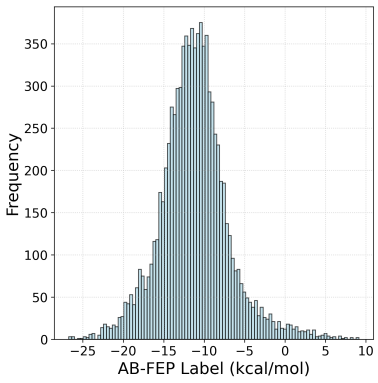

Figure 3. **Distribution of the AB-FEP calculated binding affinity labels in ToxBench**. These labels, representing binding free energies for the 8,770 ERα-ligand complexes, are in units of *kcal/mol* and span a range of approximately -26 to +9 *kcal/mol*.

binding affinity towards ERα for each of the 67 manually verified experimental values, the ensemble of AB-FEP calculations was systematically processed. An empirical, PDB-specific structural reorganization penalty was applied to the calculated free energies: +8.69 *kcal/mol* for the 1ERE structure and +6.11 *kcal/mol* for the 3ERT structure. Following these adjustments, the lowest (most favorable) free energy value across all sampled conformations for each ligand was designated as its final computational binding affinity. This workflow constitutes an initial algorithmic strategy to correlate the experimentally observed binding affinities with the predictions derived from conformational sampling and free energy calculations. As illustrated in Figure 2, this measurement yielded an RMSE of 1.754 *kcal/mol* and a Pearson correlation coefficient (R_p) of 0.692 when compared with the curated experimental data set. While the R_p value might appear modest, particularly when compared to R_p values from ML models evaluated on the ToxBench test set (Section 5), this can be primarily attributed to the significantly narrower dynamic range of binding affinities within this specific 67-compound experimental validation set. It is well-known that R_p values are sensitive to the data range, with narrow ranges often leading to lower R_p values for a similar level of absolute error. The RMSE of 1.754 *kcal/mol*, however, serves as a direct measure of absolute error. Achieving this level of accuracy validates the practical utility of our AB-FEP calculations, as it approaches the $\sim$1.0 *kcal/mol* uncertainty typically associated with experimental binding assays, thereby supporting the reliability of the ToxBench labels.

Data Statistics. In total, ToxBench contains 8,770 ERα-ligand complexes covering 699 unique ligand SMILES, each annotated with an AB-FEP-calculated binding free energy. Using a 70%/15%/15% random split and ensuring no SMILES overlap, we obtain 6,144 training data (covering 488 unique SMILES), 1,317 validation data (covering 102 unique SMILES), and 1,309 test data (covering 109 unique SMILES). The AB-FEP values span a range of approximately -26 to +9 *kcal/mol*. The overall distribution is shown in Figure 3.

By having a single-target focus, high-fidelity AB-FEP labels, and non-overlapping ligand splits, ToxBench offers a valuable testbed for developing and assessing ML models in binding affinity prediction.

4. DualBind Method

Here, we further propose a new approach, namely DualBind, for binding affinity prediction. DualBind employs a novel dual-loss strategy, which combines supervised mean squared error (MSE) loss with unsupervised denoising score matching (DSM) loss to effectively learn the binding energy function.

The Dual-Loss Framework. We represent a protein-ligand complex structure as $C = (\boldsymbol{A}, \boldsymbol{X})$, where $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ denotes the features of the n atoms in the complex, including both protein and ligand atoms, and $\boldsymbol{X} \in \mathbb{R}^{n\times 3}$ represents their atomic coordinates. We define a parameterized binding affinity prediction model as $E_\theta(\boldsymbol{A}, \boldsymbol{X}) : \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times 3} \to \mathbb{R}$, which produces a scalar energy value for a given complex structure. θ denotes the learnable parameters in the model.

Intuitively, the dual-loss framework in DualBind uses the DSM loss $\mathcal{L}_{\text{DSM}}$ to shape the energy landscape by guiding the gradient of the energy function, while the MSE loss $\mathcal{L}_{\text{MSE}}$ directly anchors predictions to known binding affinity labels. This complementary combination allows DualBind to capture both local structural variations and global binding trends effectively. An overview of our DualBind approach is illustrated in Figure 4.

MSE Loss. As shown in the upper branch of Figure 4, the MSE loss serves as a direct supervision signal for affinity prediction. Given a protein-ligand complex structure $(\boldsymbol{A}, \boldsymbol{X})$, the model predicts its binding affinity, and the error is calculated against the ground truth value. Formally, for a single data sample,

$$\mathcal{L}_{\text{MSE}} = \left(E_\theta(\boldsymbol{A}, \boldsymbol{X}) - y\right)^2, \tag{1}$$

where y represents the ground truth binding affinity, which corresponds to the AB-FEP-calculated label in ToxBench data. Intuitively, the MSE loss ensures that points, representing observed complex structures within the energy landscape, remain anchored to these affinity labels, preserving their alignment with the training data.

DSM Loss. Training neural networks with denoising is a well-established technique for enhancing model robustness and generalization (Bishop, 1995; Vincent et al., 2008; Kong

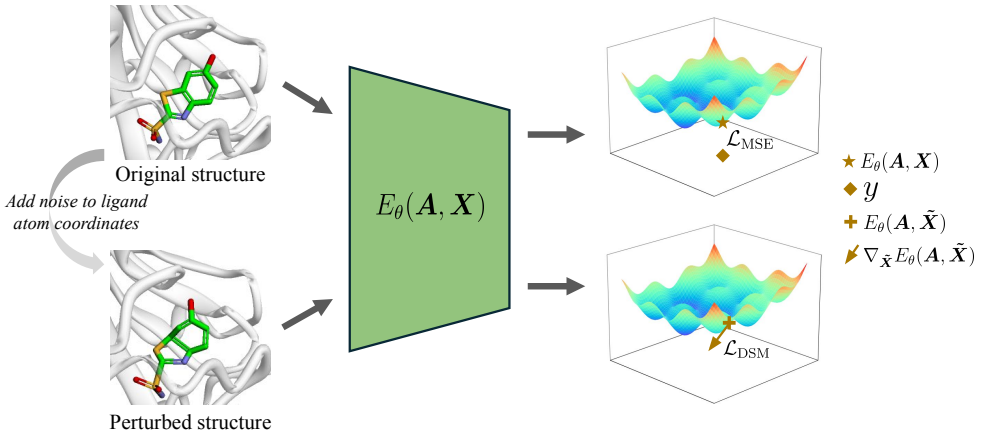

Figure 4. **An illustration of the DualBind approach.** DualBind employs a dual-loss framework that combines the MSE loss $\mathcal{L}_{\text{MSE}}$ and the DSM loss $\mathcal{L}_{\text{DSM}}$. Specifically, $\mathcal{L}_{\text{MSE}}$ anchors the predicted binding affinity of the original structure to its groundtruth label. Concurrently, $\mathcal{L}_{\text{DSM}}$ shapes the gradient of the energy function at the perturbed structure. Details are described in Section 4.

et al., 2020; Godwin et al., 2021). In particular, the denoising score matching (DSM) technique has been employed to pretrain models for 3D molecular tasks, demonstrating that such objective not only simulates learning a molecular force field but also significantly boosts performance across various downstream tasks (Zaidi et al., 2022). Furthermore, within the framework of energy-based models (EBMs), the DSM objective can be interpreted as modeling the likelihood of observed data (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021; Jin et al., 2023a). By training an EBM to denoise perturbed data, we implicitly maximize the likelihood of observing the original, unperturbed data under the modeled distribution. Thus, in the context of protein-ligand binding energy modeling, the DSM objective ensures that *the original complex structure is guided toward the local minima of the learned energy landscape*

Given such generalization capability and ability to shape local minima for observed complex structures, we incorporate the DSM objective into our DualBind model. Following previous DSM-based studies (Zaidi et al., 2022; Jin et al., 2023b), given a complex struture $(\boldsymbol{A}, \boldsymbol{X})$ from the training dataset, we first perturb it by adding Gaussian noise specifically to the ligand atom coordinates. Formally, for each ligand atom coordinate $\boldsymbol{x}_i$, the perturbed coordinate $\tilde{\boldsymbol{x}}_i$ can be obtained by

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \sigma \boldsymbol{\epsilon}_i, \quad \text{where } \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{I}_3). \tag{2}$$

σ is a hyperparameter that controls the noise scale. Note that noise is added only to ligand atoms, while protein atoms remain fixed. For simplicity, we denote the full coordinate matrix, including perturbed ligand atoms and unperturbed protein atoms, as $\tilde{\boldsymbol{X}}$.

As illustrated in the lower branch in Figure 4, the perturbed structure is fed into the model to obtain the predicted energy $E_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}})$. Then the DSM loss matches the score of the model distribution p_θ at the perturbed data $\tilde{\boldsymbol{X}}$ with the score of the distribution q where $\tilde{\boldsymbol{X}}$ is sampled. Score is defined as the gradient of the log-probability *w.r.t.* $\tilde{\boldsymbol{X}}$. Formally,

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{q(\tilde{\boldsymbol{X}})} \left[\left\| \nabla_{\tilde{\boldsymbol{X}}} \log p_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}}) - \nabla_{\tilde{\boldsymbol{X}}} \log q(\tilde{\boldsymbol{X}}) \right\|^2 \right]. \tag{3}$$

As shown by Vincent (2011), this is equivalent to

$$\begin{aligned} \mathcal{L}_{\text{DSM}} = \mathbb{E}_{q(\tilde{\boldsymbol{X}}|\boldsymbol{X}) p_{\text{data}}(\boldsymbol{X})} \Big[\Big\| & \nabla_{\tilde{\boldsymbol{X}}} \log p_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}}) \\ & - \nabla_{\tilde{\boldsymbol{X}}} \log q(\tilde{\boldsymbol{X}}|\boldsymbol{X}) \Big\|^2 \Big]. \end{aligned} \tag{4}$$

The model's energy function defines the probability distribution as $p_\theta(\boldsymbol{A}, \boldsymbol{X}) = \frac{\exp(-E_\theta(\boldsymbol{A}, \boldsymbol{X}))}{Z_\theta}$, where Z_θ is the normalizing constant, which is typically intractable but a constant *w.r.t.* $\boldsymbol{X}$ (LeCun et al., 2006; Song & Kingma, 2021). Therefore, the first term in Eq. (4) can be computed by

$$\begin{aligned} \nabla_{\tilde{\boldsymbol{X}}} \log p_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}}) &= -\nabla_{\tilde{\boldsymbol{X}}} E_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}}) - \nabla_{\tilde{\boldsymbol{X}}} \log Z_\theta \\ &= -\nabla_{\tilde{\boldsymbol{X}}} E_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}}). \end{aligned} \tag{5}$$

The second term in Eq. (4) can be easily computed as $\nabla_{\tilde{\boldsymbol{X}}} \log q(\tilde{\boldsymbol{X}}|\boldsymbol{X}) = -\frac{(\tilde{\boldsymbol{X}} - \boldsymbol{X})}{\sigma^2}$ since $q(\tilde{\boldsymbol{X}}|\boldsymbol{X})$ is a Gaussian distribution. Hence,

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{q(\tilde{\boldsymbol{X}}|\boldsymbol{X}) p_{\text{data}}(\boldsymbol{X})} \left[\left\| \nabla_{\tilde{\boldsymbol{X}}} E_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}}) - \frac{(\tilde{\boldsymbol{X}} - \boldsymbol{X})}{\sigma^2} \right\|^2 \right]. \tag{6}$$

Intuitively, minimizing this loss encourages the model to shape its energy landscape such that energy valleys (*a.k.a.*, local minima) align with the original unperturbed protein-ligand structures.

The final training loss is defined as a weighted sum of the MSE loss and the DSM loss, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{DSM}}, \tag{7}$$

where λ is a hyperparameter that balances the contribution of the two losses.

The Parameterization of $E_\theta(\boldsymbol{A}, \boldsymbol{X})$. It is important to note that our dual-loss framework is model-agnostic, allowing seamless integration with various model architectures for $E_\theta(\boldsymbol{A}, \boldsymbol{X})$. In this work, we adopt the architecture from Jin et al. (2023b), which employs an SE(3)-invariant model, built on a frame averaging neural network (Puny et al., 2021), to obtain atom representations. The key difference is that we use attention layers (Vaswani et al., 2017) within the frame averaging framework rather than the SRU++ architecture (Lei, 2021). Then, binding affinity predictions are computed by capturing pairwise atom interactions within a predefined distance threshold. For further details on this architecture, we refer readers to Jin et al. (2023b).

5. Experiments

In this section, we describe the experimental setup, present the benchmarking results, and analyze the findings.

Benchmarked Models. In addition to our proposed DualBind, we include two representative baselines in our benchmark study. Recognizing that an exhaustive evaluation of all published ML models for binding affinity prediction is impractical, our selection aims to demonstrate the performance of distinct model designs on ToxBench and establish initial baselines to facilitate future research using this dataset. The included models are:

- *Chemprop* (Heid et al., 2023). We include Chemprop, a widely-adopted model for molecular property prediction, as a ligand-only baseline. It uses a directed message passing neural network (Yang et al., 2019a) that operates solely on the 2D molecular graph of the ligand to predict the binding affinity, without incorporating any information about the protein. Its inclusion allows us to establish the performance level achievable using only ligand information within the ToxBench context, providing an important reference point for evaluating interaction-aware models.
- *AEV-PLIG* (Warren et al., 2024). AEV-PLIG is a recently proposed interaction-aware model for binding affinity prediction. AEV-PLIG first uses atomic environment vectors (AEVs) (Smith et al., 2017) to capture local atomic chemical environments derived from the 3D protein-ligand complex. These pre-computed AEV features are then processed by a graph attention network (Brody et al., 2022) to predict binding affinity. This approach, based on fixed pre-computed structural features, differs from our DualBind model, which is designed to learn relevant geometric representations directly from 3D structures.
- *DualBind*. As detailed in Section 4, our proposed DualBind employs a 3D-invariant model trained with a novel dual-loss strategy to predict binding affinity from the 3D protein-ligand complex.

Evaluation Metrics. We use the following standard regression metrics to quantitatively evaluate performance on the ToxBench task.

- *Pearson Correlation Coefficient (R_p)*. This measures the linear correlation between the predicted affinities and the ground truth affinities. Values range from -1 to 1, where 1 denotes perfect positive linear correlation, 0 indicates no linear correlation, and -1 indicates perfect negative linear correlation. Higher positive values are better.
- *Coefficient of Determination (R^2)*. Also known as R-squared, this metric represents the proportion of the variance in the ground truth affinities that is predictable from the predicted affinities. R^2 values closer to 1 indicate a better fit of the model to the data.
- *Spearman Rank Correlation Coefficient (ρ)*: This metric assesses the strength and direction of the monotonic relationship between the ranks of the predicted and true affinities. Ranging from -1 to 1, it is less sensitive to outliers than Pearson R_p and is evaluating a model's ability to correctly rank ligands according to their binding affinity, a critical aspect in virtual screening scenarios. Values closer to 1 indicate better ranking performance.
- *Root Mean Square Error (RMSE)*. RMSE measures the standard deviation of the prediction errors, providing a measure of the absolute error magnitude in the units of the target variable (*i.e.*, *kcal/mol* in the ToxBench context). Lower RMSE values indicate better predictive accuracy.

Implementation Details. We train and evaluate all three models using the predefined ToxBench splits. Final models for test set reporting are selected based on the checkpoint achieving the lowest RMSE on the validation set. Acknowledging that AB-FEP-calculated affinities greater than -3.0 *kcal/mol* typically indicate non-binding or very weak interactions, we apply a thresholding approach in our experiments. Specifically, ToxBench labels with original affinity values exceeding -3.0 *kcal/mol* are capped at -3.0 *kcal/mol* prior to training. Correspondingly, during inference, predicted affinity values greater than -3.0 *kcal/mol* are also

adjusted to -3.0 *kcal/mol* before performance metrics are computed. For robustness, all models are trained with three independent random seeds and the average performance metrics are reported.

For Chemprop and AEV-PLIG, we adapt the official open-sourced implementations and train them on the ToxBench dataset. For ChemProp, we use the default parameters in their v2.1.0 release. For AEV-PLIG, we use the optimized hyperparameters reported in their official implementations as a starting point. Key hyperparameters, such as learning rate and training epochs, are subsequently fine-tuned by monitoring the RMSE on the ToxBench validation set.

Our implementation of DualBind is based on PyTorch (Paszke et al., 2017) and PyTorch Lightning (Falcon, 2019). For protein structure input, DualBind uses a local 3D region comprising the 50 residues closest to the ligand. During training, the noise scale σ for the DSM objective, as defined in Eq. (2), is uniformly sampled from the interval [0.1, 1]. The weight λ for the DSM loss component is set to 2. We employ the Adam optimizer (Kingma & Ba, 2014) as our training optimizer. Key hyperparameters, including learning rate, batch size, and dropout rate, are tuned against the ToxBench validation set. The final configuration uses a learning rate of 5×10^{-4}, a batch size of 128, and a dropout rate of 0.1. The learning rate is decayed by a factor of 0.95 after each epoch. The model is trained for a total of 120 epochs. Training for the DualBind model is conducted on 8 NVIDIA A100 GPUs and typically completed in $\sim$4 hours. The model has a total of $\sim$1.02 million learnable parameters.

Benchmarking Results. We evaluate the trained Chemprop, AEV-PLIG, and our proposed DualBind model on the ToxBench test set using the previously defined metrics. Figure 5 visually summarizes the comparative performance, while the precise numerical results are detailed in Table 1. Our proposed DualBind model consistently outperforms both Chemprop and AEV-PLIG across all evaluation metrics. DualBind achieves the highest Pearson R_p of 0.844, the highest R^2 of 0.704, the highest Spearman ρ of 0.786, and the lowest RMSE of 2.392 *kcal/mol*. Figure 6 further provides a qualitative illustration of DualBind's strong predictive performance, presenting a scatter plot of affinities predicted by a single model (without ensembling) against the ground-truth AB-FEP-calculated values on the test set.

A particularly significant finding from our benchmark is the substantial performance gap observed between the ligand-only Chemprop model and the interaction-aware models, including DualBind and AEV-PLIG. Chemprop achieves an RMSE of 3.502 *kcal/mol*, markedly worse than those achieved by AEV-PLIG (2.645 *kcal/mol*) and DualBind (2.392 *kcal/mol*). This advantage for interaction-aware models is also clear across all other metrics, as shown in Table 1. Such observed gap is critical when contextualized with the limitations of existing sparse, multi-target datasets like PDBBind. As discussed in Section 2, models trained on such benchmarks can achieve surprisingly strong performance by exploiting ligand-only or protein-only features due to inherent dataset biases, without necessarily learning true protein-ligand interactions (Volkov et al., 2022). In contrast, the comparatively worse performance of the ligand-only model on ToxBench indicates that our dataset design, focusing on a single target with dense AB-FEP labeling, effectively mitigates such dataset-specific shortcuts. This design forces ML models to engage more deeply with the complexities of protein-ligand binding interactions to achieve high accuracy. In summary, the clear advantage demonstrated by interaction-aware models on ToxBench validates its utility as a robust benchmark for fostering the development of ML methods that truly learn the principles of molecular interaction for a specific target, a crucial step towards practical, target-centric drug discovery.

Table 1. **Performance comparison on the ToxBench benchmark.** Reported values for each method are the mean and standard deviation over three independent runs.

Method	$R_p(\uparrow)$	$R^2(\uparrow)$	$\rho(\uparrow)$	RMSE($\downarrow$)
Chemprop	$0.669_{\pm 0.011}$	$0.445_{\pm 0.016}$	$0.613_{\pm 0.014}$	$3.502_{\pm 0.051}$
AEV-PLIG	$0.790_{\pm 0.004}$	$0.616_{\pm 0.010}$	$0.756_{\pm 0.004}$	$2.645_{\pm 0.033}$
DualBind	$\mathbf{0.844}_{\pm 0.018}$	$\mathbf{0.704}_{\pm 0.034}$	$\mathbf{0.786}_{\pm 0.025}$	$\mathbf{2.392}_{\pm 0.136}$

High-Throughput Affinity Prediction with ML. In addition to assessing predictive accuracy, ToxBench is designed to unlock the potential of ML models to approximate binding affinities from high-fidelity physics-based methods like AB-FEP, but with a substantial reduction in computational cost. To be specific, a typical AB-FEP calculation for a single protein-ligand complex, such as those used to generate the labels in ToxBench, requires approximately 35 hours on an NVIDIA T4 GPU. In contrast, inference with the trained DualBind model is orders of magnitude faster. Measured over the ToxBench test set, DualBind's average inference time for a single complex on an NVIDIA A100 GPU is merely 126 *ms* without batching. This further improves to an average of 33 *ms* per complex when employing a practical batch size of 96. Remarkably, DualBind achieves a 10^6-fold speed-up compared to the typical AB-FEP calculation time. Such a dramatic increase in throughput, coupled with the competitive predictive performance as described above (RMSE of 2.392 *kcal/mol*), demonstrates the potential capability of carefully designed ML models like DualBind to approximate AB-FEP quality results at a fraction of the computational cost. This is significant for high-throughput binding affinity prediction, allowing for rapid screening of large-scale chemical libraries to accelerate target-based drug discovery.

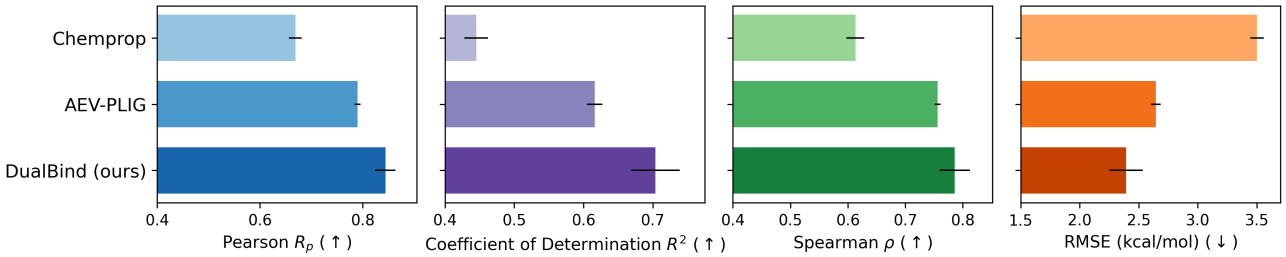

Figure 5. **Performance comparison on the ToxBench benchmark.** The evaluation metrics include R_p (Pearson correlation coefficient), R^2 (coefficient of determination), ρ (Spearman's rank correlation coefficient), and RMSE (root mean square error). ↑ (↓) represents that a higher (lower) value denotes better performance. Results presented are the averages obtained from three independent runs for each method. Detailed numerical results are provided in Table 1.

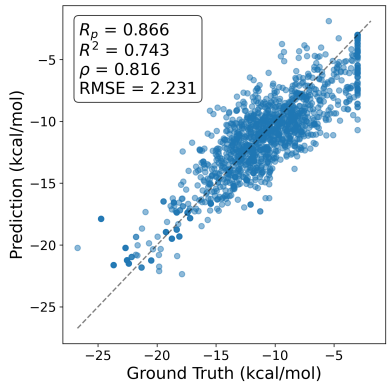

Figure 6. **DualBind predictions *vs.* ground truth**. This compares DualBind's predicted binding affinities from the best performing of three runs against the corresponding ground truth values on the ToxBench test set.

6. Conclusion and Outlook

In this work, we introduce ToxBench, a novel benchmark dataset comprising AB-FEP-calculated binding affinities specifically for Human Estrogen Receptor Alpha, a pharmaceutically critical target. ToxBench is designed with a single-target focus to facilitate the development of ML models that genuinely learn protein-ligand interactions. Our benchmarking study validates this design, demonstrating that the ligand-only model significantly underperforms interaction-aware models on ToxBench, in contrast to the observations on prior benchmarks. Furthermore, our proposed DualBind model achieves state-of-the-art performance on this new benchmark. Importantly, our experiments show that ML models have great potential to approximate the high fidelity AB-FEP calculations while offering orders of magnitude computational speed-up, thereby paving the way for high-throughput applications.

In summary, ToxBench represents a significant advance towards more robust and practically relevant ML-driven binding affinity prediction. It offers a valuable testbed for the community to develop, evaluate, and compare ML models. Moreover, the design principles behind ToxBench provide a blueprint for creating similarly impactful datasets across other key biological targets.

References

Aldeghi, M., Heifetz, A., Bodkin, M. J., Knapp, S., and Biggin, P. C. Accurate calculation of the absolute free energy of binding for drug molecules. *Chemical science*, 7(1):207–218, 2016.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Beroza, P., Crawford, J. J., Ganichkin, O., Gendelev, L., Harris, S. F., Klein, R., Miu, A., Steinbacher, S., Klingler, F.-M., and Lemmen, C. Chemical space docking enables large-scale structure-based virtual screening to discover rock1 kinase inhibitors. *Nature Communications*, 13(1): 6447, 2022.

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Böhm, H.-J. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of computer-aided molecular design*, 8:243–256, 1994.

Böhm, H.-J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3d database search programs. *Journal of computer-aided molecular design*, 12: 309–309, 1998.

Boresch, S., Tettinger, F., Leitgeb, M., and Karplus, M. Absolute binding free energies: a quantitative approach for their calculation. *The Journal of Physical Chemistry B*, 107(35):9535–9551, 2003.

Brody, S., Alon, U., and Yahav, E. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

Chen, W., Cui, D., Jerome, S., Michino, M., Lenselink, E., Huggins, D., Beautrait, A., Vendome, A., Abel, R., Friesner, R. A., and Wang, L. Enhancing hit discovery in virtual screening through absolute protein–ligand binding free-energy calculations. *J. Chem. Inf. Model.*, 63(10): 3171–3185, 2023.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., and Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11): 1046–1051, Nov 2011. doi: 10.1038/nbt.1990. URL https://doi.org/10.1038/nbt.1990.

Durant, G., Boyles, F., Birchall, K., Marsden, B., and Deane, C. M. Robustly interrogating machine learning-based scoring functions: what are they learning? *Bioinformatics*, 41(2):btaf040, 2025.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design*, 11:425–445, 1997.

Falcon, W. A. Pytorch lightning. *GitHub*, 3, 2019.

Gilson, M. K., Given, J. A., Bush, B. L., and McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical journal*, 72(3):1047–1069, 1997.

Godwin, J., Schaarschmidt, M., Gaunt, A., Sanchez-Gonzalez, A., Rubanova, Y., Veličković, P., Kirkpatrick, J., and Battaglia, P. Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971*, 2021.

Gohlke, H. and Case, D. A. Converging free energy estimates: MM-PB (GB) SA studies on the protein–protein complex Ras–Raf. *Journal of computational chemistry*, 25(2):238–250, 2004.

Gohlke, H., Hendlich, M., and Klebe, G. Predicting binding modes, binding affinities andhot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspectives in Drug Discovery and Design*, 20:115–144, 2000.

Gohlke, H., Kiel, C., and Case, D. A. Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the Ras–Raf and Ras–RalGDS complexes. *Journal of molecular biology*, 330 (4):891–913, 2003.

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., and Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.*, 47:1750–1759, 2004.

Head, R. D., Smythe, M. L., Oprea, T. I., Waller, C. L., Green, S. M., and Marshall, G. R. Validate: A new method for the receptor-based prediction of binding affinities of novel ligands. *Journal of the American Chemical society*, 118(16):3959–3969, 1996.

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. Chemprop: a machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Homeyer, N. and Gohlke, H. Free energy calculations by the molecular mechanics poisson- boltzmann surface area method. *Molecular informatics*, 31(2):114–122, 2012.

Jiang, D., Hsieh, C.-Y., Wu, Z., Kang, Y., Wang, J., Wang, E., Liao, B., Shen, C., Xu, L., Wu, J., et al. InteractionGraphNet: a novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *Journal of medicinal chemistry*, 64(24): 18209–18232, 2021.

Jiménez, J., Skalic, M., Martinez-Rosell, G., and De Fabritiis, G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling*, 58(2):287–296, 2018.

Jin, W., Chen, X., Vetticaden, A., Sarzikova, S., Raychowdhury, R., Uhler, C., and Hacohen, N. DSMBind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv*, pp. 2023–12, 2023a.

Jin, W., Sarkizova, S., Chen, X., Hacohen, N., and Uhler, C. Unsupervised protein-ligand binding energy prediction via neural euler's rotation equation. *Advances in Neural Information Processing Systems*, 36, 2023b.

Johnston, R. C., Yao, K., Kaplan, Z., Chelliah, M., Leswing, K., Seekins, S., Watts, S., Calkins, D., Chief Elk, J., Jerome, S. V., Repasky, M. P., and Shelley, J. C. Epik: pka and protonation state prediction through machine learning. *J. Chem. Theory Comput.*, 19:2380–2388, 2023.

Jorgensen, W. L., Buckner, J. K., Boudon, S., and Tirado-Rives, J. Efficient computation of absolute free energies of binding by computer simulations. application to the methane dimer in water. *The Journal of chemical physics*, 89(6):3742–3746, 1988.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research*, 33(12): 889–897, 2000.

Kong, K., Li, G., Ding, M., Wu, Z., Zhu, C., Ghanem, B., Taylor, G., and Goldstein, T. Flag: Adversarial data augmentation for graph neural networks. 2020.

La Merrill, M. A., Vandenberg, L. N., Smith, M. T., Goodson, W., Browne, P., Patisaul, H. B., Guyton, K. Z., Kortenkamp, A., Cogliano, V. J., Woodruff, T. J., et al. Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification. *Nature Reviews Endocrinology*, 16(1):45–57, 2020.

LeCun, Y., Chopra, S., and Hadsell, R. A tutorial on energy-based learning. 2006.

Lee, H.-R., Jeung, E.-B., Cho, M.-H., Kim, T.-H., Leung, P. C., and Choi, K.-C. Molecular mechanism (s) of endocrine-disrupting chemicals and their potent oestrogenicity in diverse cells and tissues that express oestrogen receptors. *Journal of cellular and molecular medicine*, 17(1):1–11, 2013.

Lei, T. When attention meets fast recurrence: Training language models with reduced compute. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7633–7648, 2021.

Li, J., Guan, X., Zhang, O., Sun, K., Wang, Y., Bagni, D., and Head-Gordon, T. Leak proof PDBBind: A reorganized dataset of protein-ligand complexes for more generalizable binding affinity prediction. *ArXiv*, pp. arXiv–2308, 2024.

Liu, X., Jiang, S., Duan, X., Vasan, A., Liu, C., Tien, C.-c., Ma, H., Brettin, T., Xia, F., Foster, I. T., et al. Binding affinity prediction: From conventional to machine learning-based approaches. *arXiv preprint arXiv:2410.00709*, 2024.

Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., and Wang, R. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50(2):302–309, 2017.

Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., and Zheng, S. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*, 35:7236–7249, 2022.

Meli, R., Morris, G. M., and Biggin, P. C. Scoring functions for protein-ligand binding affinity prediction using structure-based deep learning: a review. *Frontiers in bioinformatics*, 2:885983, 2022.

Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7 (2):146–157, 2011.

Miller, E. B., Murphy, R. B., Sindhikara, D., Borrelli, K. W., Grisewood, M. J., Ranalli, F., Dixon, S. L., Jerome, S., Boyles, N. A., Day, T., Ghanakota, P., Mondal, S., Rafi, S. B., Troast, D. M., Abel, R., and Friesner, R. A. Reliable and accurate solution to the induced fit docking problem for protein–ligand binding. *J. Chem. Theory Comput.*, 17(4):2630–2639, 2021. doi: 10.1021/acs.jctc.1c00136. URL https://doi.org/10.1021/acs.jctc.1c00136.

Miziak, P., Baran, M., Błaszczak, E., Przybyszewska-Podstawka, A., Kałafut, J., Smok-Kalwat, J., Dmoszyńska-Graniczka, M., Kiełbus, M., and Stepulak, A. Estrogen receptor signaling in breast cancer. *Cancers*, 15(19):4689, 2023.

Moon, S., Zhung, W., Yang, S., Lim, J., and Kim, W. Y. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, 13(13):3661–3673, 2022.

Murphy, R. B., Repasky, M. P., Greenwood, J. R., Tubert-Brohman, I., Jerome, S. S., Annabhimoju, R., Boyles, N. A., Schmitz, C. D., Abel, R., Farid, R., and Friesner, R. A. WScore: A flexible and accurate treatment of explicit water molecules in ligand–receptor docking. *J. Med. Chem.*, 59(9):4364–4384, 2016. doi: 10.1021/acs.jmedchem.6b00113.

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, jul 2012. doi: 10.1021/jm300687e. URL https://doi.org/10.1021/jm300687e.

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., and Venkatesh, S. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37 (8):1140–1147, 2021.

Öztürk, H., Özgür, A., and Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Pagadala, N. S., Syed, K., and Tuszynski, J. Software for molecular docking: a review. *Biophysical reviews*, 9(2): 91–102, 2017.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Puny, O., Atzmon, M., Smith, E. J., Misra, I., Grover, A., Ben-Hamu, H., and Lipman, Y. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2021.

Ross, G., Lu, C., Scarabelli, G., and others. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry*, 6(1):222, oct 2023. doi: 10.1038/s42004-023-01019-9. URL https://doi.org/10.1038/s42004-023-01019-9.

Sadybekov, A. V. and Katritch, V. Computational approaches streamlining drug discovery. *Nature*, 616(7958): 673–685, 2023.

Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aid. Mol. Des.*, 27(3):221–234, 2013. doi: 10.1007/s10822-013-9644-8.

Scantlebury, J., Vost, L., Carbery, A., Hadfield, T. E., Turnbull, O. M., Brown, N., Chenthamarakshan, V., Das, P., Grosjean, H., Von Delft, F., et al. A small step toward generalizability: training a machine learning scoring function for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 63(10):2960–2974, 2023.

Schrödinger Release 2021-3: FEP+. Schrödinger, LLC, New York, NY, 2021. Software release.

Schrödinger, LLC. Schrödinger Release 2024-4: LigPrep, 2024.

Smith, J. S., Isayev, O., and Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science*, 8(4): 3192–3203, 2017.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y. and Kingma, D. P. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428, 2021.

Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society*, 112(16):6127–6129, 1990.

Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. Comparative assessment of scoring functions: the CASF-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2018.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Volkov, M., Turk, J.-A., Drizard, N., Martin, N., Hoffmann, B., Gaston-Mathé, Y., and Rognan, D. On the frustration to predict binding affinities from protein–ligand structures with deep neural networks. *Journal of medicinal chemistry*, 65(11):7946–7958, 2022.

Wang, J. and Dokholyan, N. V. Yuel: improving the generalizability of structure-free compound–protein interaction prediction. *Journal of chemical information and modeling*, 62(3):463–471, 2022.

Wang, R., Lai, L., and Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design*, 16:11–26, 2002.

Wang, R., Fang, X., Lu, Y., and Wang, S. The PDBbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

Warren, M., Deane, C., Magarkar, A., Morris, G., Biggin, P., et al. How to make machine learning scoring functions competitive with fep. 2024.

Yang, J., Shen, C., and Huang, N. Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. *Frontiers in pharmacology*, 11:69, 2020.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019a.

Yang, X., Wang, Y., Byrne, R., Schneider, G., and Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical reviews*, 119(18):10520–10594, 2019b.

Yuan, W., Chen, G., and Chen, C. Y.-C. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Briefings in Bioinformatics*, 23(1):bbab506, 2022.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., and Leach, A. R. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, January 2024. doi: 10.1093/nar/gkad1004.