DualBind: A Dual-Loss Framework for Protein-Ligand Binding Affinity Prediction

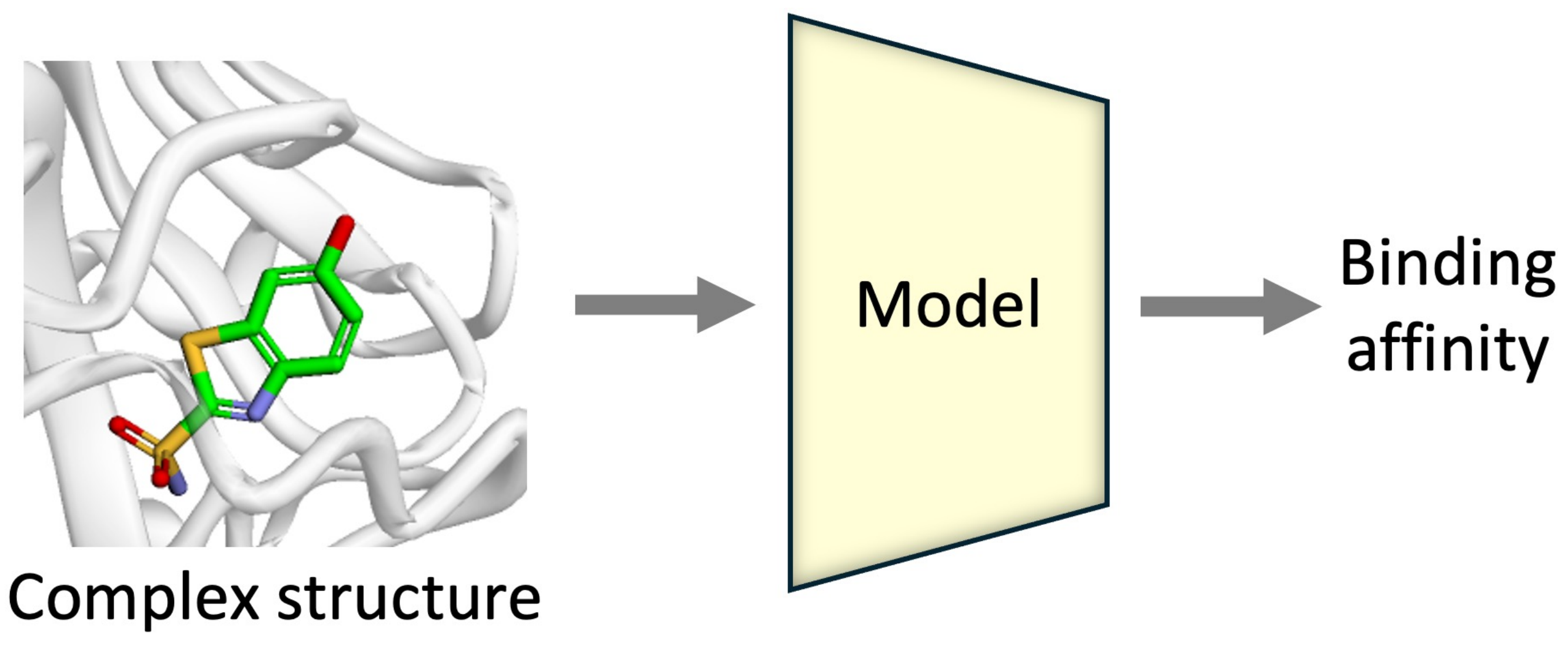
Meng Liu, Saee Gopal Paliwal

TLDR

We present DualBind, a simple and effective dual-loss framework that integrates supervised mean squared error (MSE) with unsupervised denoising score matching (DSM) for accurate binding affinity prediction.

INTRODUCTION

Binding affinity prediction is fundament for drug discovery.


Complex structure

Model

Binding affinity

An illustration of the binding affinity prediction task

Supervised approaches

- Require reliable binding affinity labels
- Easy to overfit on limited data

DSMBind [1] adopts a generative modeling strategy by training an energy-based model (EBM) with a denoising score matching (DSM) objective.

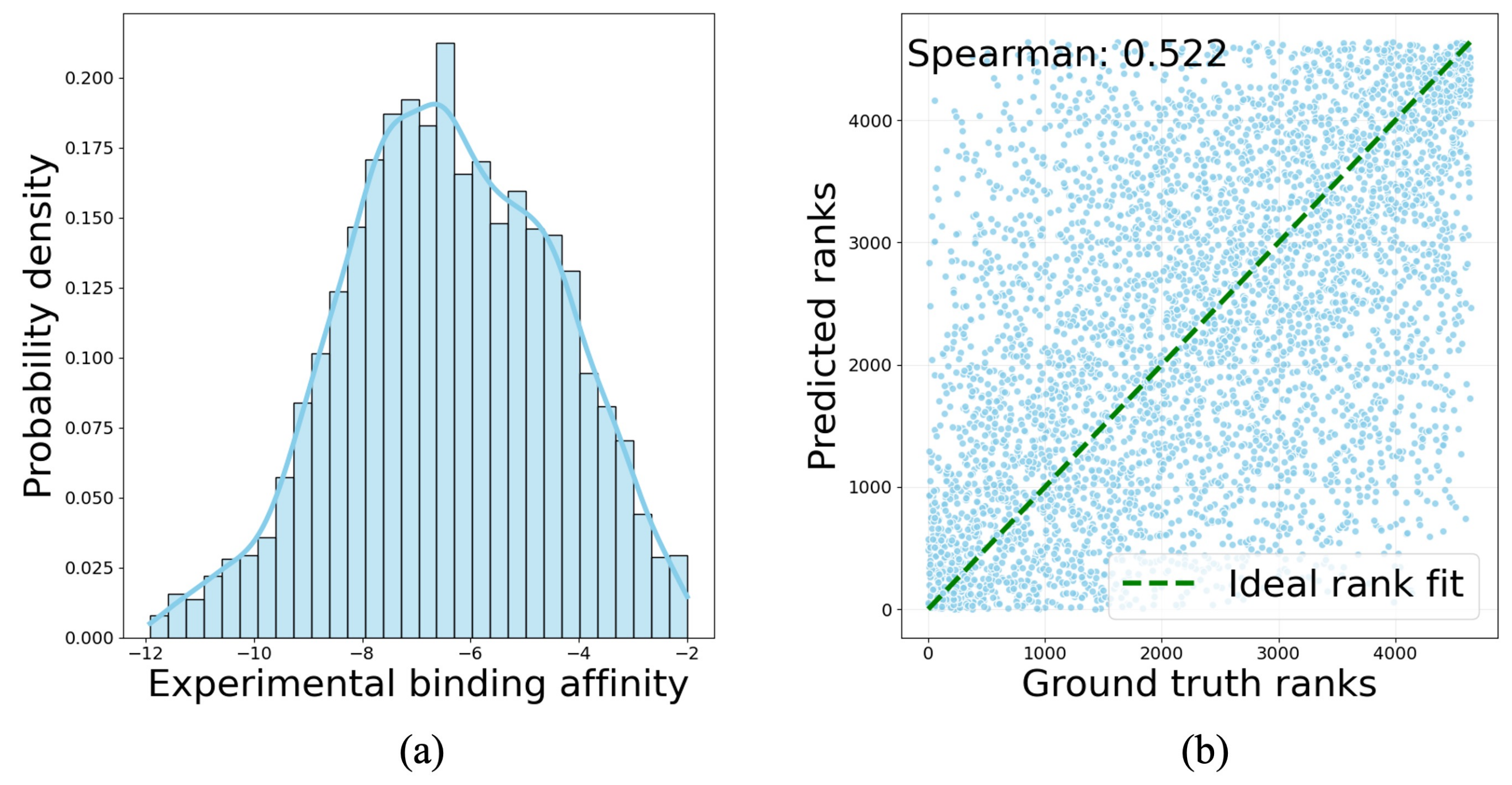
- Maximize the likelihood of training structures, without requiring binding affinity labels
- Cannot produce absolute affinities, but the learned energy function **correlates** with binding energies

METHODOLOGY

Boltzmann distribution assumption in DSM models: The effectiveness of the DSM objective, which aims to **precisely learn the energy function by maximizing data likelihood**, depends on the assumption that training samples follow a Boltzmann distribution, $P(C) \propto e^{-E(C)}$.

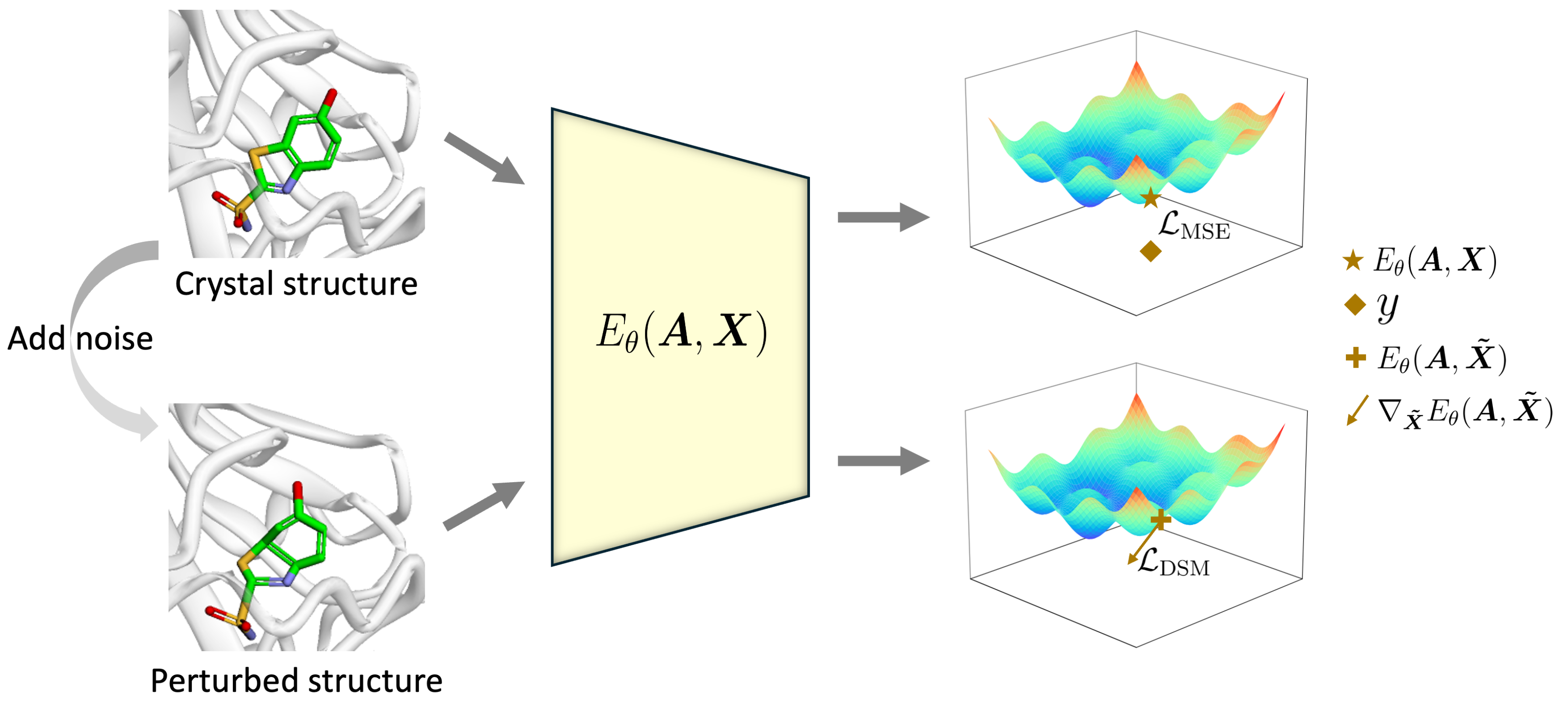
The actual distribution of complexes in experimental datasets often diverges from this assumption due to experimental biases, selective data reporting, *etc.*

Thus, although the DSM objective can effectively assign local minima (gradient is zero) to observed protein-ligand complexes, we conjecture the learned function struggles to accurately rank their **relative binding affinities**.



(a) Distribution of binding affinity in the PDBbind v2020 refined dataset.
(b) Rank fit of a DSM-only model on training complexes.

DualBind is a dual-loss framework combines the DSM loss $\mathcal{L}_{DSM}$, which learns the energy landscape by **shaping the gradient** of the energy function, with the MSE loss $\mathcal{L}_{MSE}$, which directly **ties the predictions to known binding affinity values**.


Crystal structure

Add noise

Perturbed structure

$E_\theta(\boldsymbol{A}, \boldsymbol{X})$

$\mathcal{L}_{\text{MSE}}$

$\mathcal{L}_{\text{DSM}}$

★ $E_\theta(\boldsymbol{A}, \boldsymbol{X})$

◆ y

✚ $E_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}})$

↙ $\nabla_{\tilde{\boldsymbol{X}}} E_\theta(\boldsymbol{A}, \tilde{\boldsymbol{X}})$

An illustration of the DualBind methodology

$\mathcal{L}_{DSM}$ shapes the gradient of the energy landscape such that the **energy valleys (local minima) align with the unperturbed crystal structures**.

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

Advantages

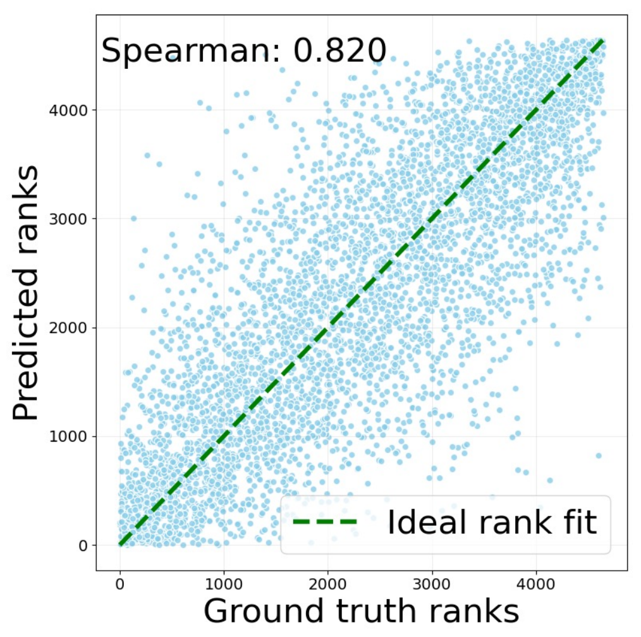
- Produce more accurate **absolute** affinity predictions, rather than merely **comparative** values provided by DSM-only models
- Exhibit better **generalization** capability compared to MSE-only models because of the denoising technique.
- Has the unique capability to harness the full potential of **both labeled and unlabeled data**. The dual-loss framework allows to utilize both labeled and unlabeled data for training by calculating their corresponding loss values.

EXPERIMENTS

Benchmark results demonstrate the above advantages.

Method	Affinity labels	R_p↑	RMSE↓	ρ↑
Glide-XP	✗	0.467	1.95	-
Glide-SP	✗	0.513	1.89	-
Autodock Vina	✗	0.604	1.73	-
DSMBind (Gaussian)	✗	0.638	N/A	-
DSMBind (SE(3))	✗	0.656	N/A	-
K_{DEEP}	✓	0.738	1.462	-
PIGNet	✓	0.749	-	-
DualBind	✓	**0.757**±0.006	**1.461**±0.013	**0.742**±0.008
MSE-only	✓	0.749±0.008	1.491±0.017	0.736±0.006
DSM-only	✗	0.646±0.005	N/A	0.652±0.007

Performance comparison on the CASF-2016 benchmark



Rank fit of DualBind on training complexes

Preliminary experiment shows unique ability of DualBind to utilize both labeled data and unlabeled data.

Method	#Labeled	#Unlabeled	R_p↑	RMSE↓	ρ↑
MSE-only	2321	✗	0.664±0.037	1.694±0.086	0.666±0.028
DualBind	2321	2321	**0.731**±0.007	**1.684**±0.087	**0.732**±0.006
MSE-only	4643	✗	0.749±0.008	1.491±0.017	0.736±0.006

Experimental results on DualBind's flexible data use strategy

[1] Wengong Jin, Siranush Sarkizova, Xun Chen, Nir Hacohen, and Caroline Uhler. "Unsupervised protein-ligand binding energy prediction via neural euler's rotation equation." Advances in Neural Information Processing Systems 36 (2024).