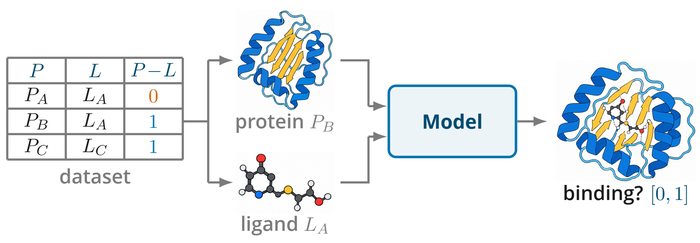
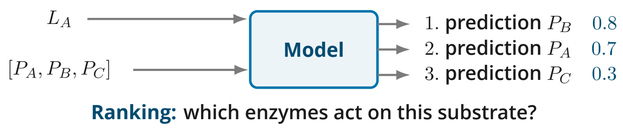
When Accuracy Misleads: Ligand-Conditional Ranking in Enzyme–Substrate Interaction Benchmarks

Global AUC can be high while the model ranks each ligand's true target close to chance. Ranking metrics catch this. Four training components fix it.

Christoph Manitz[1,2], Georg Künze[1,2,3]
[1]Institute for Drug Discovery, [2]ScaDS.AI, [3]Interdisciplinary Centre for Bioinformatics (IZBI), Leipzig University, Germany

1. THE PROBLEM





Ranking: which enzymes act on this substrate?

THE METRICS

Global AUC: true pair scored above a false pair, all ligands pooled.

$$\frac{TP + TN}{\text{all predictions}}$$

Per-ligand AUC: the same within each ligand, then averaged.

$$\frac{\mathrm{AUC}_{L1} + \cdots + \mathrm{AUC}_{Ln}}{n_L}$$

MRR: 1/rank of each ligand's true protein, averaged.

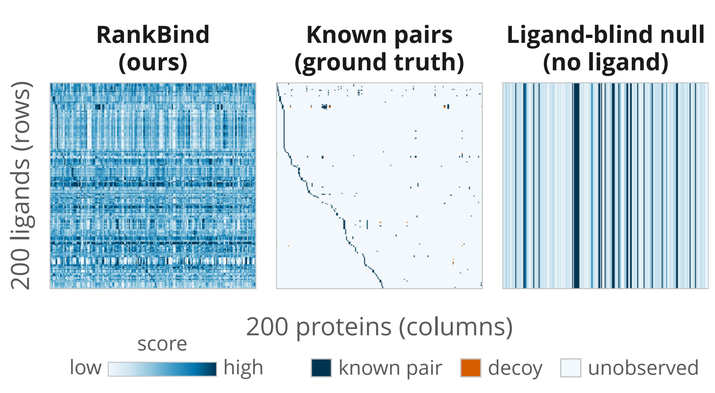
$$\mathrm{mean}\left(\frac{1}{\mathrm{rank}_p}\right) \text{ (per ligand)}$$

Hit@10: share of true proteins in their ligand's top 10.

$$\mathrm{Hit@10} = \frac{\#\text{top 10 hits}}{\#\text{true pairs}}$$

Model	Global AUC	per-lig. AUC	MRR	H@10	prior-top share	Gini
DrugBAN	0.955	0.551	0.048	0.103	0.30	0.991
MolTrans	0.937	0.455	0.046	0.082	0.44	0.992
GraphDTA	0.827	0.589	0.093	0.196	0.97	0.990
ligand-blind null	0.500	—	—	—	1.00	0.995
RankBind	0.620	**0.825**	**0.305**	**0.574**	**0.13**	**0.843**

Ranking metrics: each test ligand vs. all 132 test proteins (chance: AUC 0.5, MRR 0.042, prior-top share 0.13). RankBind: 10 seeds.



The shortcut, made visible. A stripe is a protein winning almost every ligand. Middle: the 376 documented pairs (orange: decoys). The GraphDTA baseline is in column 3.

TAKE HOME

All three models achieve a high Global AUC (0.83–0.95), but this does not translate into ligand-specific ranking: per-ligand AUCs stay close to chance (0.45–0.59).

Global AUC certifies dataset-level regularities, not ligand-conditional discrimination.

2. THE FIX

For each ligand, rank its candidate proteins. Do not learn one global protein-popularity score.

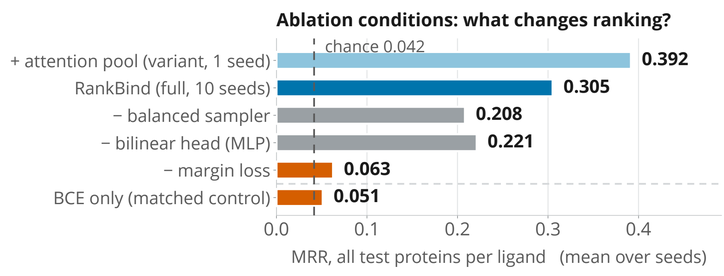
COMPONENTS

Within-ligand margin loss. Compare candidate proteins for one ligand and push the true target above the others.

Online hard-negative mining. Each epoch, draw the negatives from the non-substrate proteins the current model scores highest for that ligand, instead of at random.

Protein-balanced sampling. Keep positive and negative pairs approximately balanced for each protein, so frequent proteins cannot dominate batches.

Low-rank bilinear head. Let ligand and protein representations interact directly instead of concatenating them in an MLP.

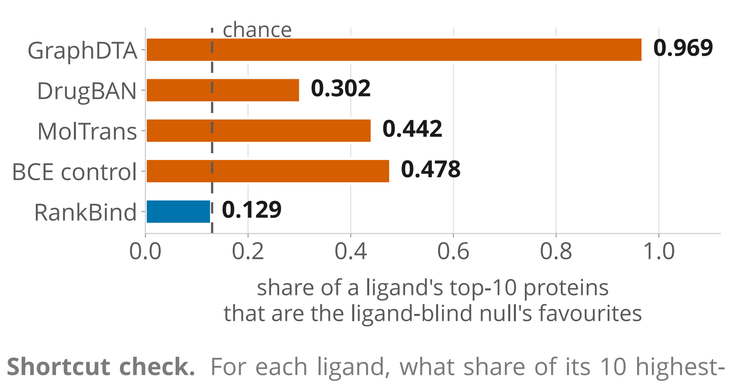


Ablation study. "–" rows remove one component. "BCE only": matched control with the same encoders; it learns each pair's 0/1 label independently with binary cross-entropy. Every test ligand ranked against all 132 test proteins; 3 seeds per ablation row, 10 for the full model (seed s.d. 0.111); chance MRR 0.042. Attention-pool variant: one seed; the default on that seed reaches 0.391.

TAKE HOME

MRR rose from 0.051 (control, 3 seeds) to 0.305 (10 seeds; chance 0.042); Global AUC traded down from 0.918 to 0.620.

Ranking improves because the training design changes, not because Global AUC improves.



Shortcut check. For each ligand, what share of its 10 highest-scored proteins are among the 26 proteins most often labelled positive in training? A ranking that ignores that prior scores 0.13.

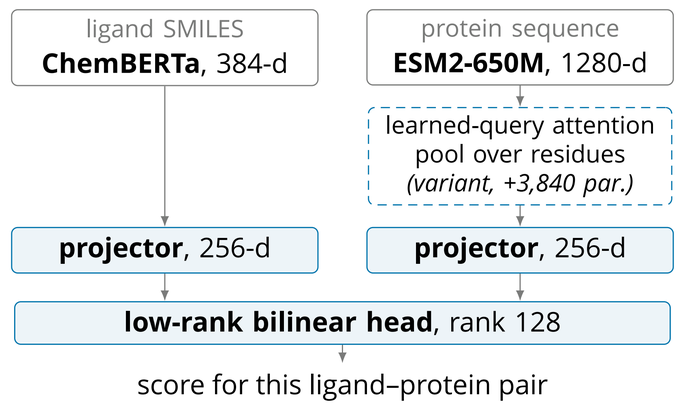
DOES IT ONLY MEMORISE?

MRR (chance 0.042); identity to nearest training protein.

Test subset	RankBind	control
156 unseen ligands	**0.172**	0.038
proteins $<$ 30% identity	**0.119**	0.019
proteins $\geq$ 70% identity	0.473	0.121

Part of the signal is recognition, but not all of it.

3. WHAT IS TRAINED



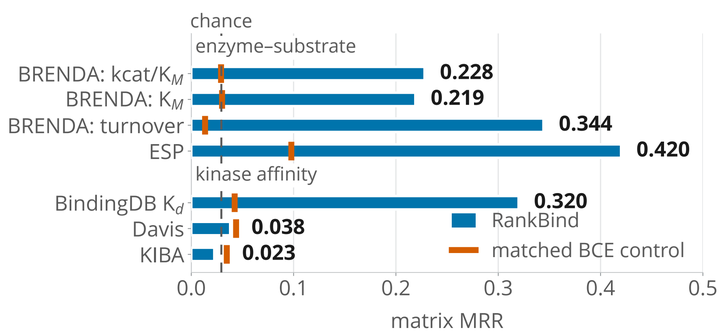
Blue: trained (627,201 parameters). Grey: frozen.

Deliberately simple: training, not architecture, fixes the shortcut. Only the projectors and the head are trained.

4. DOES IT GENERALISE?

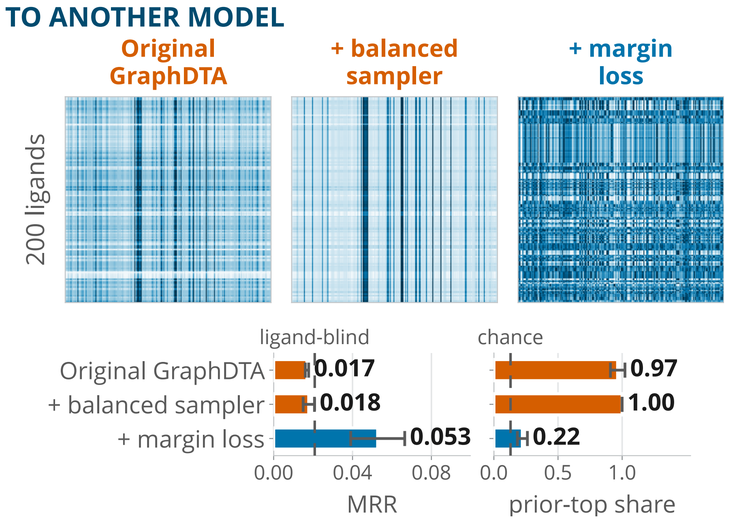
TO OTHER DATASETS

7 datasets, same components. The matched BCE control passes **Global AUC 0.84–0.98** at near-chance MRR 0.014–0.098. RankBind ranks **7–25**× better on the kinetic sets.



RankBind beats the control on every enzyme-substrate set. ESP (Kroll et al. 2023), independent: pool 4,000, AUC 0.894.

TO ANOTHER MODEL



The ranking loss transfers; the encoder still matters. Dashed: ligand-blind prior, chance. RankBind, same pool: MRR 0.247.

TAKE HOME

RankBind beats the control on 5/7 datasets, by 7–25× on the kinetic ones.

The ranking loss alone moves GraphDTA off the shortcut: the recipe transfers.

Report Global AUC only after ranking per ligand.

scads.ai

christoph manitz.de

RankBind GitHub

kuenzelab.org

CONTACT
Leipzig University
Christoph Manitz
christoph.manitz@uni-leipzig.de

IN COOPERATION WITH
UNIVERSITÄT LEIPZIG
Dresden University of Technology

FUNDED BY
Federal Ministry of Research, Technology and Space
SACHSEN

Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.