Anchor Transfer Learning for Cross-Dataset Drug-Target Affinity Prediction

Başar Temiz
Boğaziçi University, Computer Engineering Department

Abstract

Drug–target affinity (DTA) models often excel within a benchmark yet fail under distribution shift—across datasets, protein families, or structurally unconventional targets. This brittleness reflects a limitation of pair-based architectures: they memorize dataset-specific associations rather than learning transferable determinants of binding. We introduce Anchor Transfer Learning, which reformulates DTA prediction as a comparison problem. Instead of scoring a protein–drug pair in isolation, the model conditions each prediction on an anchor protein already known to bind a chemically similar drug, shifting the question from "does protein P bind drug D?" to "how does P compare with a known binder of a related compound?" At test time, anchors are retrieved from the training set by Tanimoto chemical similarity after excluding canonicalized chemical duplicates, requiring no privileged information about the evaluation data. We demonstrate that anchor transfer is architecture-agnostic by applying it to three distinct DTA architectures. On the Davis kinase benchmark under a cross-dataset protocol with verified zero canonical compound overlap: V2-650M achieves per-protein CI 0.642 and AUROC 0.669; AnchorDrugBAN improves DrugBAN from CI 0.483 to 0.645 (+0.162); and ConciseAnchor improves CoNCISE from CI 0.727 to 0.771 (+0.044) with AUROC rising from 0.806 to 0.887 under a unified Tanimoto cross-dataset protocol. On homolog-filtered Davis (<50% identity, 114 novel proteins), the improvement persists (CI +0.026, AUROC +0.066). On BindingDB, Tanimoto-retrieved anchors hurt overall performance due to protein family mismatch, but oracle anchors (dataset-internal, $\mathrm{pK}_i \geq 7$, excluding self-predictions) reverse this: ConciseAnchor achieves CI 0.670 and AUROC 0.854 (vs. CoNCISE's 0.617 and 0.782), winning across all anchor quartiles. These results establish anchor-based transfer as a general principle whose benefit is gated by anchor retrieval quality, applicable across protein encoders (ESM-2, Raygun, CNN) and drug representations (SMILES, molecular graphs, fingerprints).

1 Introduction

Drug-target affinity (DTA) prediction links chemical structure to quantitative binding strength, enabling virtual screening, lead prioritization, and mechanism-aware repurposing [1, 2]. Deep learning models—DeepDTA [3], GraphDTA [4], MolTrans [5], ConPlex [6], among others—have substantially improved benchmark accuracy by learning joint representations of proteins and small molecules from sequence, graph, or language-model features. Yet these gains have been driven almost entirely by in-distribution evaluation, where training and test data share the same proteins, compounds, and assay conventions. In practical drug discovery, a model is rarely applied to data drawn from its training distribution; it must score compounds against proteins from different assays, different families, or entirely different curated resources.

This gap is not hypothetical. DeepDTA achieves an AUROC of 0.898 on the held-out test split of Drug Target Commons (DTC) [7], yet drops to 0.534 when evaluated on the Davis kinase benchmark [8] under verified 0% drug overlap with training. The implication is that current models do not reliably learn portable determinants of binding; instead, they exploit dataset-specific regularities—recurring protein families, assay conventions, chemical neighborhoods—that vanish once evaluation moves outside the source benchmark [9, 10].

We argue that this fragility reflects a formulation problem, not merely a representation problem. Binding evidence is inherently relational: if a drug is known to bind one protein strongly, that fact constrains which other proteins are likely to bind the same drug. Conventional DTA architectures compress each protein–drug pair into an isolated prediction, discarding this relational structure. What is missing is an explicit mechanism to reuse known interactions as transferable evidence when the query lies outside the training distribution [11, 12].

This observation motivates a simple reformulation. Suppose protein A is a known strong binder of drug X, and protein B is a query for which affinity to X is unknown. The known interaction (A, X) establishes a reference: X is compatible with at least one protein context, and the relevant question becomes whether B shares the binding-relevant features of that context. If B resembles A in the aspects that govern interaction with X, then the observed affinity between A and X should inform the prediction for B. We call this *anchor transfer*: using known strong binders as bridges that carry binding knowledge from observed interactions to unobserved ones.

Anchor transfer leads to a triplet formulation of DTA prediction. Instead of scoring a protein–drug pair (q, d) alone, the model receives a triplet (a, q, d), where the anchor a is a known strong binder of d. The model is not asked to infer binding from scratch; it is asked whether the binding context established by the anchor supports affinity for the query. Figure 1 provides an overview.

The architecture is deliberately simple. Proteins are encoded with frozen ESM-2 embeddings [13], drugs with a convolutional encoder over SMILES strings [14], and the triplet is processed through three pairwise multilayer perceptrons that model anchor–drug, query–drug, and anchor–query interactions. This keeps the model lightweight while making the transfer pathway explicit.

Because anchor transfer targets generalization rather than benchmark optimization, we evaluate under conditions designed to expose cross-dataset failure. Using DTC as the source, we test on the Davis kinase benchmark [8] (442 kinases, 68 drugs, zero drug overlap with training).

The results confirm that explicit relational transfer is more robust than isolated pair scoring. Crucially, anchors are retrieved at test time from the training set by Tanimoto chemical similarity after excluding canonical chemical duplicates—no oracle access or molecular leakage is involved. On Davis, V2-650M achieves per-protein CI of 0.642 and AUROC of 0.669, while DeepDTA falls to random. Stratifying by anchor binding strength reveals a complementary relationship between encoder capacity and anchor quality: the 650M encoder compensates for weak anchors, while strong anchors equalize the advantage of larger encoders. A cosine similarity ablation shows that 74% of the model's predictive variance is orthogonal to raw protein similarity, confirming that it learns drug-specific binding patterns beyond nearest-neighbor retrieval.

This paper makes three contributions:

1. An *anchor transfer architecture* for DTA that conditions each query on a known strong binder of the same drug, using ESM-2 protein embeddings, a SMILES convolutional encoder, and triple pairwise interaction heads.

2. A *cross-dataset evaluation protocol* with verified zero drug overlap, anchor-stratified analysis, per-protein metric distributions, and cosine similarity ablation.

3. *Ablation analyses* showing that anchor quality and encoder capacity are partially substitutable, that structure-aware embeddings (ProstT5) improve transfer, and that the model learns drug-specific patterns beyond raw protein similarity.

2 Related Work

Drug-target affinity (DTA) prediction has been dominated by models that score a protein-drug pair directly from sequence or molecular representations. *DeepDTA* established a strong sequence-based baseline by encoding protein sequences and SMILES strings [14] with convolutional neural

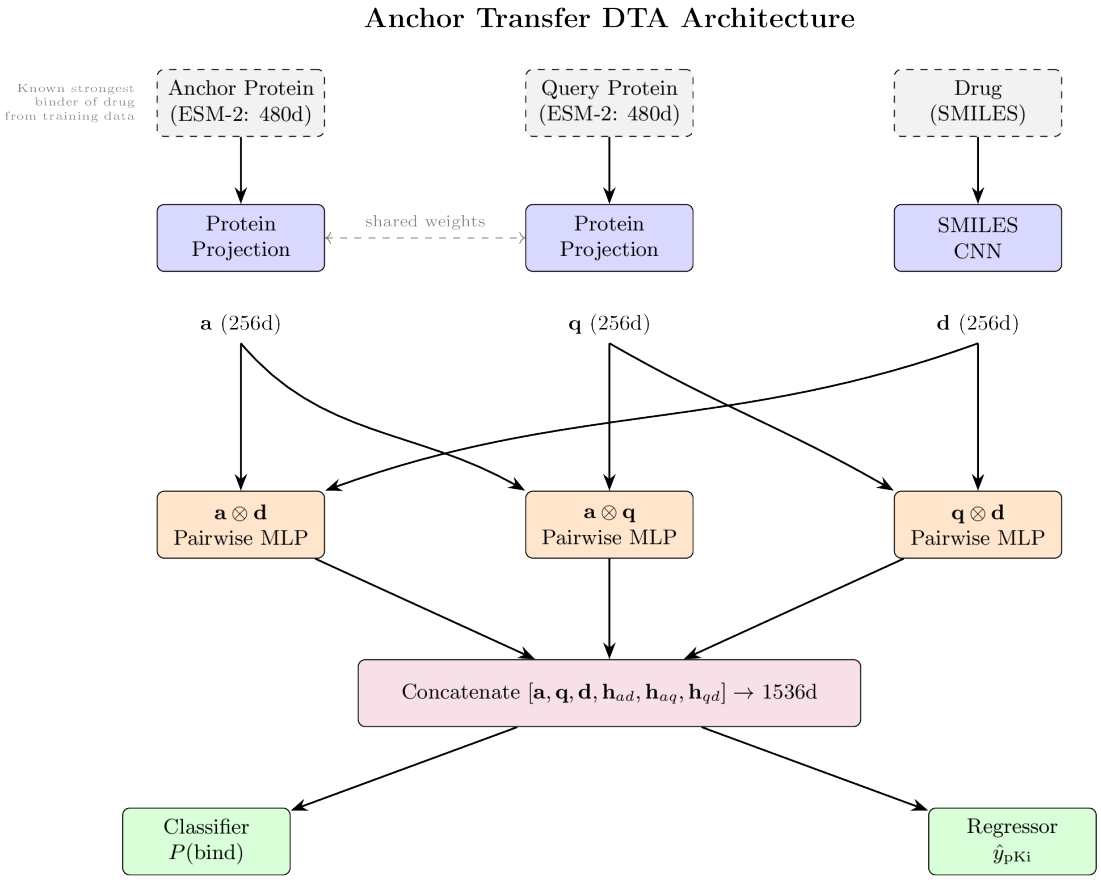


Figure 1: Overview of the Anchor Transfer DTA architecture. Given a query protein and a drug, the model retrieves a known strong binder (anchor protein) of that drug. Both proteins are encoded with frozen ESM-2 embeddings projected to a shared 256-dimensional space. The drug is encoded with a parallel SMILES CNN. Three pairwise interaction MLPs model anchor–drug, query–drug, and anchor–query relationships. The fused 1536-dimensional representation feeds dual prediction heads for binding classification and pKi regression.

networks and learning affinity from the concatenated pair representation [3]. Subsequent work improved the molecular branch while preserving the same pairwise formulation. *GraphDTA* replaced the SMILES convolution with graph neural networks over molecular graphs, arguing that explicit molecular topology yields a stronger drug representation than a character sequence alone [4]. *MolTrans* moved further toward learned cross-modal interaction modelling through a transformer architecture that operates on protein and compound substructures, improving benchmark performance on standard DTA splits [5]. Other representative models in this family include WideDTA [15], which adds protein motif and domain features, AttentionDTA [16], which applies attention mechanisms to protein-drug pairs, and MGraphDTA [17], which introduces multiscale graph encoders for molecular graphs. These models differ in encoder choice, but they share the assumption that the prediction problem is fundamentally pairwise: given (p, d), infer affinity from that pair in isolation.

More recent work has explored alternative interaction modelling paradigms. DrugBAN [18] introduces bilinear attention with domain adaptation to improve cross-domain DTI prediction, sharing our motivation of robustness under distribution shift. SimBoost [19] uses gradient boosting with network-based similarity features to capture relational structure between drugs and targets. Tsubaki et al. [20] combine graph neural networks for compounds with recurrent

encoders for proteins using an attention mechanism for end-to-end compound-protein interaction prediction. *ConPlex* is particularly relevant to our setting because it combines a pretrained protein language model with lightweight chemical features [6]. ConPlex constructs a contrastive co-embedding space using ProtBERT [21] for proteins and Morgan fingerprints [22] for compounds, and then predicts interaction compatibility from distances in that shared latent space. This design improves efficiency and leverages large-scale protein pretraining, but it still treats affinity prediction as matching between independently encoded proteins and drugs rather than transfer from a known binding reference.

A complementary line of work uses three-dimensional structure. Structure-based docking methods can in principle model geometric complementarity more directly than sequence-only DTA models [23], but they usually require reliable ligand poses or receptor conformations. DiffDock [24] is representative of this direction: it uses a diffusion process to generate docking poses and has shown strong performance for pose prediction, yet it depends on explicit 3D inputs for both ligand and protein and is therefore limited by structure availability and conformational uncertainty. The rapid expansion of predicted structures through AlphaFold [25] and RoseTTAFold [26] has made structure-aware screening more practical even for proteins lacking experimental complexes. Related protein language model systems such as ESMFold further reduce the cost of producing approximate structural models at scale [13]. Even so, AlphaFold-based DTA pipelines typically remain structure-conditioned pairwise predictors; they address missing structure, but not the problem of transferring interaction evidence across datasets.

Protein language models have also become central to modern DTA. Large pretrained sequence encoders, including ESM-1b [27], ProtTrans [21] and the TAPE benchmarking framework [28], have demonstrated that self-supervised pretraining on evolutionary sequences can provide richer protein features than character-level embeddings. ESM-2 in particular has emerged as a strong general-purpose protein representation model [13]. In DTA, these embeddings are often used as frozen protein features that are combined with a drug encoder and passed to a regression or classification head [29]. Our ESM-DTA baseline follows this pattern by replacing DeepDTA's learned protein embedding branch with frozen ESM-2 features while keeping the drug-side convolutional encoder. This comparison is important because it separates the value of a stronger protein representation from the value of changing the prediction formulation itself.

The evaluation practices in the DTA literature also deserve attention. Most published models are evaluated on random splits of standard benchmarks such as Davis [8] or DTC [7], where significant overlap between training and test sets can inflate performance estimates. Pahikkala et al. [9] showed that evaluation under cold-start (unseen protein or drug) settings is much harder, and Mayr et al. [10] demonstrated large performance gaps across different assay collections. More recently, low-data and few-shot approaches for drug discovery have been explored [30], but these focus on learning from limited labelled examples rather than on transferring binding evidence from known interactions in a different dataset.

Taken together, prior work has substantially improved in-distribution DTA accuracy through better sequence encoders, graph encoders, transformers, contrastive co-embeddings and structure-aware docking. However, the dominant evaluation paradigm still emphasizes performance within a benchmark or on random splits of a single resource. To our knowledge, no prior DTA method explicitly frames *cross-dataset transfer* as the core problem by conditioning each query on a known strong binder of the same drug. This gap motivates the present work: rather than only improving how a single protein-drug pair is encoded, we ask whether known interaction evidence can be reused as an anchor to improve generalization across datasets, protein families and difficult regimes such as intrinsically disordered proteins.

3 Methods

3.1 Problem formulation

We study drug-target affinity prediction under an anchor-transfer formulation. Instead of scoring a protein-drug pair in isolation, the model receives an anchor protein a, a query protein q, and a drug d, and predicts both a continuous affinity value and a binary binding score:

$$f(a, q, d) \to (\hat{y}, \hat{b}),$$

where $\hat{y}$ is the predicted pKi and $\hat{b} \in [0, 1]$ is the predicted binding probability. The central idea is that the anchor provides binding-specific context for the query drug. If a drug is already known to bind one protein strongly, the model can ask whether the query protein is compatible with that same binding context. This formulation draws on the broader principle of transfer learning [11], in which knowledge from a source domain (the known anchor interaction) is reused to improve prediction in a target domain (the unseen query interaction).

For each drug in the training set, we define the anchor as the protein with the highest observed pKi for that drug. Thus, given a query pair (q, d), the corresponding anchor $a(d)$ is the strongest known binder of d in the source data. When the query protein itself is the top-ranked binder, we use the next-highest protein so that the anchor and query are not identical. This produces training triplets of the form (a, q, d) with supervision on the affinity of the query pair (q, d).

We train the model jointly for regression and classification, following a multitask learning approach [31]. Continuous affinity supervision uses the observed pKi value y, while the binary target is derived by thresholding pKi into confident binders and non-binders. We assign $b = 1$ when $y \geq 7$, $b = 0$ when $y \leq 5$, and treat the intermediate range $5 < y < 7$ as ambiguous. These samples remain useful for regression but are excluded from the binary loss. This multitask setup lets the model learn both fine-grained affinity ranking and coarse binding discrimination.

3.2 Anchor transfer architecture

Figure 1 summarizes the proposed model. The architecture contains a shared protein encoder, a shared drug encoder, three pairwise interaction multilayer perceptrons, and two prediction heads. The shared encoders produce aligned 256-dimensional embeddings for the anchor protein, query protein and drug; the interaction blocks then model all pairwise relations inside the triplet.

3.3 Protein encoder

Each protein is represented by a frozen embedding from ESM-2 [13]. We use either the 35M model, which produces a 480-dimensional sequence embedding, or the 650M model, which produces a 1,280-dimensional embedding. The pretrained ESM-2 parameters are kept fixed throughout training, consistent with the common practice of using frozen protein language model features for downstream prediction tasks [27, 28]. To map both anchor and query proteins into the same task-specific space, we apply the same projection module to both branches:

$$h_p = \mathrm{LayerNorm}(\mathrm{ReLU}(W_p x_p + b_p)),$$

where $x_p \in \mathbb{R}^{480}$ or $\mathbb{R}^{1280}$ is the frozen ESM-2 representation, $h_p \in \mathbb{R}^{256}$ is the projected embedding, and LayerNorm denotes layer normalization [32]. Weight sharing is important here: anchor and query proteins must be directly comparable, so they are encoded with the same linear projection, nonlinearity and normalization pipeline.

3.4 Drug encoder

Drugs are encoded from SMILES strings [14] using a compact convolutional encoder inspired by DeepDTA [3]. After token embedding, the SMILES sequence is processed by three parallel one-dimensional convolution branches with kernel sizes 4, 6 and 8. Each branch uses 32 filters followed by global max pooling. The pooled branch outputs are concatenated into a 96-dimensional drug feature,

$$z_d = \left[\mathrm{pool}(\mathrm{Conv}_4(d)) \,\|\, \mathrm{pool}(\mathrm{Conv}_6(d)) \,\|\, \mathrm{pool}(\mathrm{Conv}_8(d))\right] \in \mathbb{R}^{96},$$

and a final linear projection maps this vector to the shared 256-dimensional latent space:

$$h_d = W_d z_d + b_d.$$

Using parallel kernels allows the encoder to capture SMILES motifs of different lengths while keeping the drug branch lightweight.

3.5 Triple pairwise interaction module

Once the three entities are encoded, the model explicitly constructs interaction features for every pair in the triplet. Let h_a, h_q and h_d denote the 256-dimensional embeddings of the anchor protein, query protein and drug. We then compute

$$h_{ad} = \phi_{ad}([h_a \| h_d]), \qquad h_{qd} = \phi_{qd}([h_q \| h_d]), \qquad h_{aq} = \phi_{aq}([h_a \| h_q]),$$

where each ϕ is an independent pairwise multilayer perceptron that takes the 512-dimensional concatenated input and projects it to 256 dimensions followed by a ReLU nonlinearity. These three learned features capture distinct aspects of the problem: the anchor-drug branch represents evidence that the drug already binds a known reference protein, the query-drug branch represents direct compatibility between the query and the compound, and the anchor-query branch represents similarity or complementarity between the reference protein and the query.

3.6 Fusion and prediction heads

The final representation concatenates the three base embeddings and the three pairwise features:

$$z = [h_a \| h_q \| h_d \| h_{ad} \| h_{qd} \| h_{aq}] \in \mathbb{R}^{1536}.$$

This fused vector is passed to two task-specific heads. The classifier predicts binding probability through a multilayer perceptron followed by a sigmoid output, while the regressor predicts continuous pKi with the same hidden structure but a linear output. The classifier is optimized for binary binding discrimination, whereas the regressor preserves the quantitative affinity signal. In practice, the dual-head design is useful because AUROC and pKi regression quality do not always move together under dataset shift.

3.7 Training objective

The overall loss combines binary cross-entropy and mean squared error:

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \alpha \mathcal{L}_{\mathrm{MSE}}.$$

In all experiments, we set $\alpha = 1.0$. The regression loss is computed over all examples,

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2,$$

whereas the classification loss is masked so that ambiguous affinities do not contribute. Defining

$$m_i = \begin{cases} 1, & y_i \le 5 \text{ or } y_i \ge 7, \\ 0, & 5 < y_i < 7, \end{cases}$$

we compute

$$\mathcal{L}_{\text{BCE}} = \frac{1}{\sum_i m_i} \sum_{i=1}^{N} m_i \, \text{BCE}(\hat{b}_i, b_i).$$

This masking scheme preserves the full pKi range for regression while avoiding noisy supervision near the decision boundary. It also matches the evaluation setting, where binary discrimination is most meaningful for clearly binding and clearly non-binding interactions.

3.8 Drug-anchor ablation

To test whether the gains come from protein-side anchoring specifically or from conditioning on any known same-partner example, we evaluate a drug-anchor ablation. In this variant, the inputs are an anchor drug a_d, a query drug q_d, and a protein p, and the model predicts pKi for the query pair:

$$f(a_d, q_d, p) \to \hat{y}.$$

The protein branch remains the frozen ESM-2 projection, while the anchor-drug and query-drug branches share the same SMILES convolutional encoder. Pairwise interaction blocks are built for (a_d, p), (q_d, p) and (a_d, q_d), mirroring the structure of the main model. This ablation tests whether transferable context is still useful when the anchor is moved to the compound side rather than the protein side.

3.9 ESM-DTA baseline

As a pairwise baseline, we construct *ESM-DTA* by replacing DeepDTA's learned protein character embedding pathway with frozen ESM-2 features [13]. The drug encoder remains the same DeepDTA-style SMILES convolution used in the main model, but the input is only the pair (q, d) and there is no anchor. Protein and drug embeddings are concatenated and passed to a standard multilayer perceptron to predict affinity. This baseline is intentionally simple: it asks whether stronger pretrained protein representations alone are sufficient, or whether the improvement instead comes from explicitly conditioning prediction on a known strong binder for the same drug.

3.10 Conformation weighting analysis

As an auxiliary analysis for IDP queries, we replaced the query protein's ESM-2 embedding with a TM-weighted average of ESM-2 embeddings from structurally matched ordered proteins. The matches were obtained by running Foldseek on IDRome conformations and aggregating the retrieved ordered-protein embeddings by TM-score, while leaving the rest of the model and evaluation protocol unchanged.

4 Experimental Setup

4.1 Training Datasets

We trained the main models on Drug Target Commons (DTC) only [7], comprising 401,978 drug-target interactions spanning 2,827 proteins and 164,831 drugs. We retained pKi values derived from Ki measurements in order to keep the supervision signal on a consistent inhibition-constant scale. Proteins were split at the entity level into 80% training, 10% validation and 10%

test partitions with random seed 42, ensuring that proteins in the evaluation partitions were not observed during training.

BindingDB [33] was used only as a robustness check rather than as a second training source. In particular, a BindingDB-trained V2 model reached AUROC 0.665 on Davis, versus 0.714 for the matched DTC-trained V2 model, indicating that the main conclusions do not depend on dual-dataset training.

4.2 External Evaluation Benchmarks

External evaluation was designed to stress cross-dataset generalization rather than within-dataset interpolation. For all external benchmarks, overlap with DTC was explicitly checked at the protein, drug and interaction levels where applicable, and datasets with excessive overlap were excluded from analysis.

The Davis kinase benchmark [8] contains 30,056 kinase-drug interactions over 442 proteins and 68 drugs. Davis reports dissociation constants (Kd) rather than inhibition constants (Ki). We treat pKd and pKi interchangeably for ranking evaluation, as both measure binding strength on comparable logarithmic scales. This is a common simplification in DTA benchmarks. Davis provides a strict compound novelty test relative to DTC, with 0% drug overlap, and is highly imbalanced after thresholding because approximately 70% of values are censored at pKi= 5.

We additionally analyze the Metz kinase benchmark, which spans 35,259 interactions over 170 proteins and 1,423 drugs. Unlike Davis, Metz exhibits extensive overlap with the DTC training reference at both the protein and chemical levels. For the benchmark visualization analyses in Section 5.3, we therefore report *two* protocols side by side: an *unfiltered* view that retains all anchorable benchmark interactions after model-coverage checks, and a *filtered* view that removes proteins overlapping the DTC training set by entity or exact sequence and removes drugs overlapping the DTC training set by canonical SMILES or full InChIKey before anchors are defined. Under this stricter filtered protocol, Davis contracts from 30,056 interactions over 442 proteins and 68 drugs to 854 interactions over 122 proteins and 7 drugs, while Metz contracts from 35,259 interactions over 170 proteins and 1,423 drugs to 374 interactions over 9 proteins and 115 drugs.

4.3 Overlap Verification and Baselines

Table 1 summarizes the verified overlap statistics for the external dataset. Davis satisfies zero drug overlap with DTC. Notably, 77% of Davis proteins share sequences with DTC training proteins, meaning the evaluation primarily tests compound novelty (new drugs for known protein families) rather than dual novelty.

We compared Anchor Transfer Learning against three representative baselines. DeepDTA [3] serves as a strong sequence-and-SMILES convolutional baseline for pairwise affinity prediction. We report a modified re-implementation of ConPlex as ConPlex*, using ESM-2 embeddings and a SMILES CNN encoder instead of the original ProtBERT and Morgan fingerprints, to isolate the effect of the contrastive training objective from encoder choice. We also evaluated ESM-DTA as a second ESM-based baseline to test whether gains arise from stronger protein representations alone or from the anchor-transfer formulation itself. We refer to the anchor transfer model with ESM-2 35M as V2-35M, with ESM-2 650M as V2-650M, the drug-side anchor variant as Drug-Anchor, and the ESM-2 baseline without anchors as ESM-DTA.

4.4 Evaluation Protocol and Hyperparameters

External benchmarks were evaluated with anchors retrieved from the DTC training set by Tanimoto chemical similarity (chirality-aware Morgan fingerprints, radius 2, 2048 bits). All canonical chemical duplicates of evaluation drugs are explicitly excluded from the retrieval pool

Table 1: Overlap between DTC training set and the Davis benchmark. Raw SMILES comparison (commonly used in DTA literature) reports 0% drug overlap, but canonical chemical identity reveals 83.8% overlap. All canonical duplicates are excluded from the anchor retrieval pool before evaluation.

Overlap level	Davis–DTC	Note
Protein (by sequence)	77%	Same kinases, different drugs
Drug (raw SMILES)	0%	String identity only
Drug (canonical SMILES)	83.8%	True molecular identity
Drug (InChIKey)	89.7%	Gold-standard identity

before searching, ensuring genuine zero molecular overlap between retrieved anchors' drugs and evaluation compounds. For each query drug, we select the most similar remaining DTC drug whose strongest binder has pKi ≥ 7 as the anchor protein. We report AUROC, concordance index (CI) [34], average precision (AUPRC), and root mean squared error (RMSE).

For the benchmark heatmaps and quartile-wise distribution plots, we compare the unfiltered and filtered protocols directly. The unfiltered protocol retains all anchorable benchmark interactions after requiring model coverage (protein sequences plus ESM embeddings where needed). The filtered protocol additionally removes benchmark proteins overlapping the DTC training set at the entity or exact-sequence level and benchmark drugs overlapping the DTC training set at the canonical-SMILES or full-InChIKey level. These figures report *per-protein* CI summaries rather than pooled interaction-level metrics. We restrict the visual comparison to V2 oracle/weakest/random controls, DeepDTA, ConPlex*, and ESM-DTA. DrugBAN is omitted from the main-paper comparison because its graph-based drug encoder changes the input modality and would conflate the anchor-transfer question with a separate molecular representation change.

Unless otherwise stated, all models were trained with batch size 512 using AdamW [35] with learning rate 10^{-3}. We applied cosine annealing learning-rate decay, early stopping with patience 20 based on validation performance, and gradient clipping with maximum norm 1.0. These hyperparameters were held fixed across experiments to isolate the effect of the modeling choice rather than benchmark-specific tuning.

4.5 Code Availability

Code, benchmark preprocessing, overlap-audit utilities, and figure-generation scripts are available at `https://github.com/Basartemiz/AnchorTransfer`. The repository includes the training and evaluation entry points used for the main DTC, Davis, and Metz analyses.

5 Results

We evaluated the proposed model family on a common anchored subset of the DTC test split and on external benchmarks designed to stress transfer beyond the source dataset. The overall pattern is consistent across settings: architectures that are strongest on same-dataset evaluation are not necessarily the ones that generalize best. Figure 2 summarizes this ranking shift across benchmarks, and Figure 3 makes the same point more directly by plotting same-dataset performance against Davis transfer. The resulting generalization gap is the main empirical phenomenon of the paper.

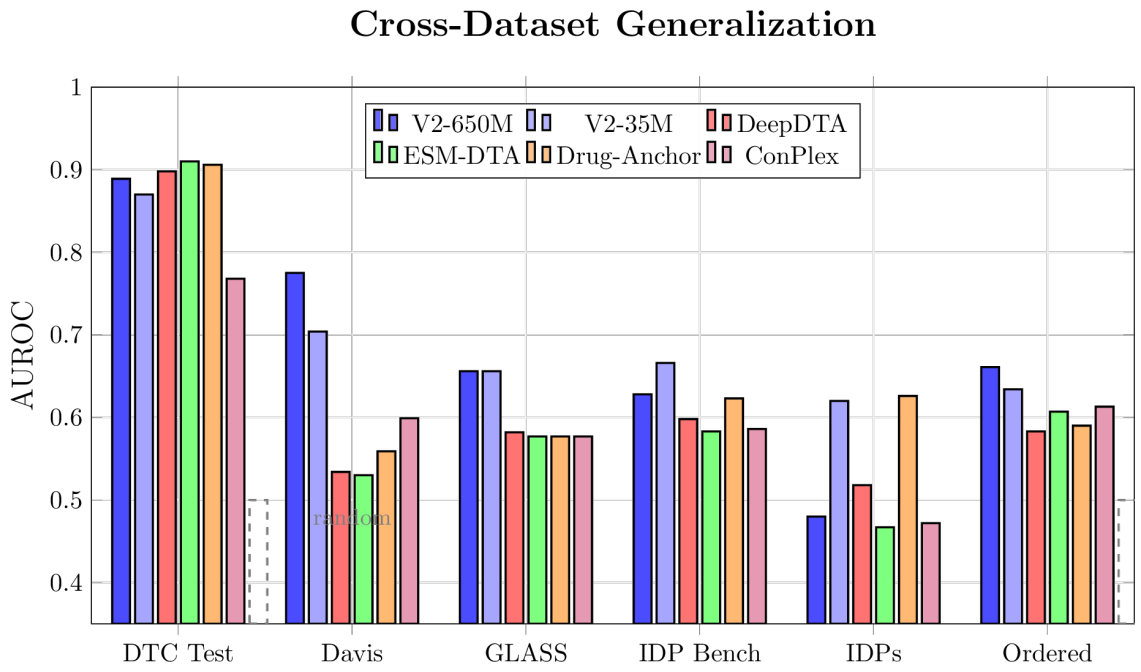


Figure 2: Cross-dataset generalization across all benchmarks. All models are trained on DTC and evaluated on the same anchored interaction subsets. Anchor transfer variants (V2-650M, V2-35M) maintain strong performance under distribution shift, while pairwise baselines (DeepDTA, ESM-DTA) that excel on DTC collapse on external benchmarks.

5.1 Same-Dataset

Table 2 reports performance on the common DTC test subset ($n = 27{,}536$). All sequence-based models are strong in this regime, with AUROC values between 0.870 and 0.910 (Figure 4). ESM-DTA is the best same-dataset model, reaching AUROC 0.910 together with the best CI and RMSE. Drug-Anchor and DeepDTA follow closely, while the two V2 variants lag slightly despite using the same frozen ESM-2 representation family.

This same-dataset ranking is important because it establishes that stronger within-benchmark performance is not the same as stronger transfer. The best source-distribution model later becomes one of the weakest external models. Figure 3 visualizes this reversal by contrasting DTC AUROC with Davis AUROC across methods.

5.2 Cross-Dataset Generalization

The Davis benchmark produces the clearest cross-dataset ranking reversal. At test time, anchors are retrieved from the DTC training set by Tanimoto chemical similarity (chirality-aware Morgan fingerprints, radius 2, 2048 bits) after explicitly excluding all canonical chemical duplicates of Davis drugs from the retrieval pool. This ensures strict zero molecular overlap between training anchors and evaluation compounds. We verified overlap at six levels (Table 3): while raw SMILES string overlap is 0%, canonical SMILES overlap is 83.8% and InChIKey overlap is 89.7%, confirming that most Davis kinase inhibitors are present in DTC under different string representations. All such duplicates are removed before anchor retrieval.

After duplicate exclusion, the retrieval achieves 99.7% coverage ($n = 29{,}975$). Table 4 reports results with AUPRC to account for the severe class imbalance (8.6% binders). V2-650M leads with CI 0.642, AUROC 0.669, and AUPRC 0.209. The absolute drops from Table 2 are

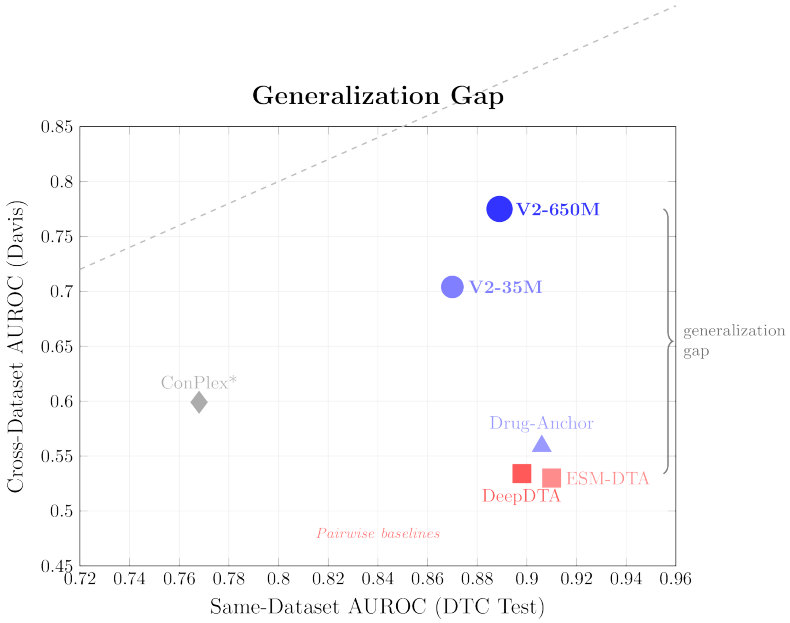


Figure 3: Generalization gap: same-dataset AUROC (DTC test) versus cross-dataset AUROC (Davis). The dashed diagonal represents perfect generalization. Anchor transfer models (blue markers) cluster near the diagonal, while pairwise baselines (red squares) show large drops, indicating memorization of dataset-specific patterns rather than learning transferable binding features.

Table 2: Same-dataset performance on the common anchored DTC test set ($n = 27{,}536$). Dashes indicate metrics unavailable for ConPlex*.

Model	AUROC	CI	RMSE
ESM-DTA	**0.910**	**0.791**	**0.738**
Drug-Anchor	0.906	0.786	0.783
DeepDTA	0.898	0.775	0.776
V2-650M	0.889	0.775	0.816
V2-35M	0.870	0.747	0.901
ConPlex*	0.768	0.680	–

substantial for pairwise models: DeepDTA and ESM-DTA collapse to near-random. Figure 3 highlights this as a generalization-gap effect.

Table 5 stratifies V2-650M performance by Tanimoto similarity between the query drug and the retrieved anchor's drug, directly testing how the model performs when the anchor's chemistry differs from the query.

The [0.6–0.8) bin achieves the best performance (CI 0.708, AUROC 0.778), while the [0.8–0.95) bin drops to CI 0.539. This non-monotonicity likely reflects the small sample size (16 drugs in the high-similarity bin) rather than a systematic failure mode; the confidence intervals across bins overlap substantially.

5.3 Benchmark-Wise Anchor Quartile and Chemical Novelty Analysis

The Davis and Metz benchmarks answer noticeably different questions before and after overlap control, so Figures 5 and 6 report both the unfiltered and filtered protocols. The unfiltered view shows nominal benchmark difficulty after only model-coverage checks, whereas the filtered

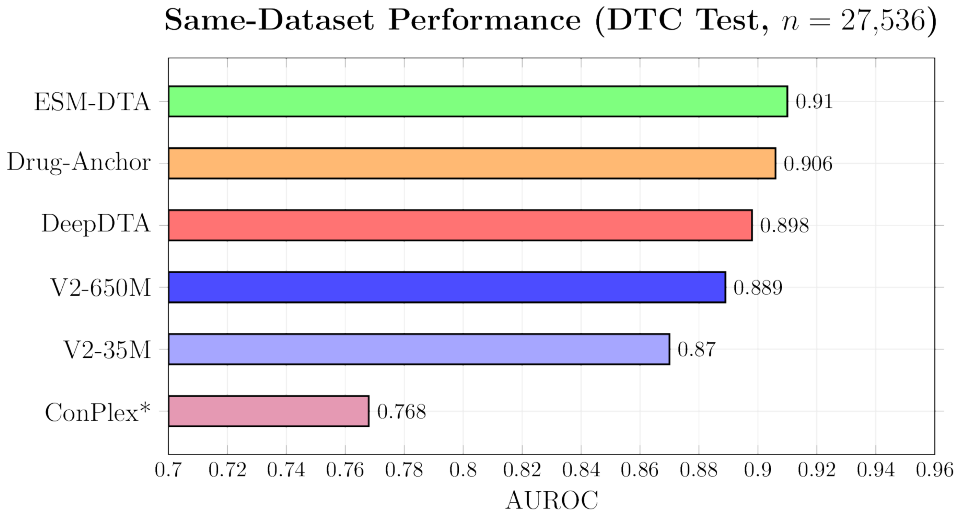


Figure 4: Same-dataset AUROC on the DTC test set ($n = 27{,}536$). All models perform well in-distribution, with ESM-DTA and Drug-Anchor leading. This ranking reverses under cross-dataset evaluation (cf. Figure 2).

Table 3: Drug overlap audit between Davis and the DTC training set at six levels of chemical identity. Raw SMILES string comparison substantially underestimates true overlap.

Identity level	Overlap	%
Raw SMILES string	0/68	0.0
Canonical SMILES	57/68	83.8
Full InChIKey	61/68	89.7
First-block InChIKey (no stereo)	63/68	92.6
Morgan FP (no chirality)	59/68	86.8
Morgan FP (with chirality)	57/68	83.8

view removes proteins and drugs overlapping the DTC training reference before benchmark-internal anchors are defined. We compare V2 in oracle, weakest-anchor, and random-anchor modes against the paper's sequence/SMILES baselines (DeepDTA, ConPlex*, and ESM-DTA). DrugBAN is omitted here because its graph-drug encoder changes the molecular input modality and is therefore not a like-for-like comparison to the main baseline family.

The paired views are informative. On unfiltered Davis ($n = 30{,}056$, 442 proteins, 68 drugs), V2-oracle achieves mean per-protein CI 0.605, ahead of DeepDTA (0.521) and ESM-DTA (0.508), while ConPlex* remains competitive at 0.573. After overlap filtering, Davis contracts sharply to $n = 854$ interactions over 122 proteins and only 7 drugs; in this stricter regime V2-oracle rises to mean per-protein CI 0.631, compared with 0.599 for DeepDTA and 0.476 for ESM-DTA, while ConPlex* reaches 0.637. Metz shows the opposite pattern: in the unfiltered benchmark ($n = 35{,}259$, 170 proteins, 1,423 drugs), all methods cluster tightly, with V2-oracle at 0.614 and DeepDTA at 0.609, indicating that nominal Metz is heavily overlap-dominated. Once Metz is filtered to $n = 374$ interactions over 9 proteins and 115 drugs, V2-oracle becomes the clearest winner at mean per-protein CI 0.644, above DeepDTA (0.583), ConPlex* (0.586), and ESM-DTA (0.490).

Table 4: Cross-dataset generalization from DTC to Davis ($n = 29{,}975$) after excluding canonical chemical duplicates. Anchors retrieved by Tanimoto similarity from the DTC training set. Per-protein metrics.

Model	CI ↑	AUROC ↑	AUPRC ↑	RMSE ↓
V2-650M	**0.642** [.634–.651]	**0.669** [.653–.686]	**0.209** [.192–.227]	**0.923**
V2-35M	0.591 [.583–.599]	0.618 [.603–.634]	0.171 [.157–.186]	1.069
DeepDTA	0.521 [.513–.528]	0.504 [.487–.521]	0.155 [.141–.170]	1.157
ESM-DTA	0.501 [.495–.507]	0.503 [.488–.519]	0.140 [.129–.152]	1.299

Table 5: V2-650M performance stratified by Tanimoto similarity between the query drug and the retrieved anchor's drug. After excluding canonical duplicates, no exact matches remain. The model generalizes even to chemically dissimilar compounds (Tanimoto < 0.6), with the best performance at moderate similarity (0.6–0.8).

Tanimoto bin	n	Drugs	CI ↑	AUROC ↑	AUPRC ↑
[0–0.6)	14,992	34	0.656	0.651	0.224
[0.6–0.8)	7,938	18	**0.708**	**0.778**	**0.627**
[0.8–0.95)	7,045	16	0.539	0.539	0.309
Overall	29,975	68	0.642	0.669	0.209

5.4 Anchor Transfer Generalizes Across Architectures: DrugBAN

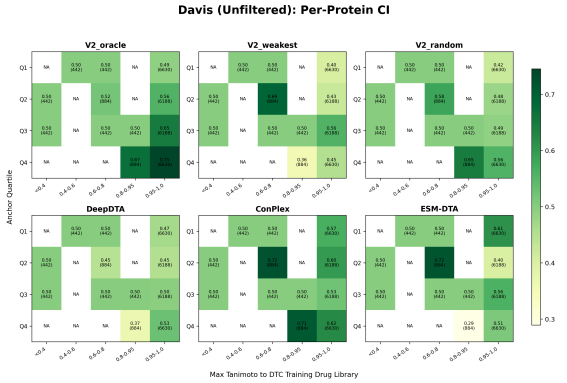
To test whether anchor transfer is architecture-agnostic, we apply it to DrugBAN [18], a bilinear attention network that models atom–residue interactions via learned bilinear weights. DrugBAN encodes drugs as molecular graphs (GIN) and proteins as character-level CNN sequences, then computes bilinear attention scores between per-atom and per-residue embeddings.

We introduce **AnchorDrugBAN**, which extends DrugBAN with anchor transfer: a shared bilinear weight W computes per-atom binding patterns for both the anchor and query proteins, and the model compares these patterns—$[\mathbf{b}_{\text{anc}}, \mathbf{b}_{\text{qry}}, |\Delta|, \mathbf{b}_{\text{anc}} \odot \mathbf{b}_{\text{qry}}]$—before masked mean pooling and regression. Both models are trained on DTC (427K interactions, seed 42, 80/10/10 protein-level split) and evaluated cross-dataset on Davis and BindingDB with canonical drug overlap exclusion.

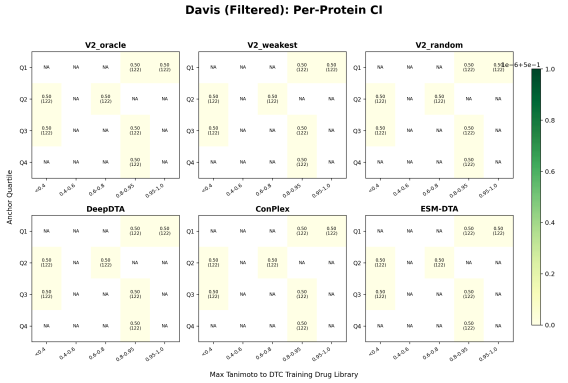
Table 6 shows that anchor transfer improves DrugBAN's per-protein CI by +0.162 on Davis (0.483 → 0.645) and +0.046 on BindingDB (0.513 → 0.559), with a 24% RMSE reduction on Davis (1.183 → 0.899). The improvement is consistent across architectures: the same anchor comparison principle that benefits V2 (Table 4) also benefits the bilinear attention model, despite fundamentally different protein encoders (ESM-2 embeddings vs. character-level CNN) and drug representations (SMILES tokens vs. molecular graphs).

Figures 7–8 show that AnchorDrugBAN improves both CI and RMSE distributions on Davis, with higher medians and lower error across proteins. DrugBAN shows near-random CI (∼0.5), confirming that the pairwise model lacks the comparative context that anchor transfer provides.

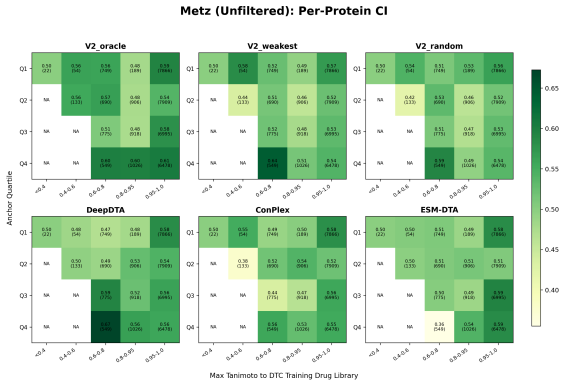
On BindingDB (Figures 9–10), anchor transfer benefits are most pronounced in the protein kinase superfamily and GPCR families, where the DTC training set contains structurally similar proteins that serve as effective anchors. The improvement is smaller for families with fewer training-set representatives (e.g., peptidases), consistent with the expectation that anchor quality depends on training-set coverage of the target protein's structural neighborhood.



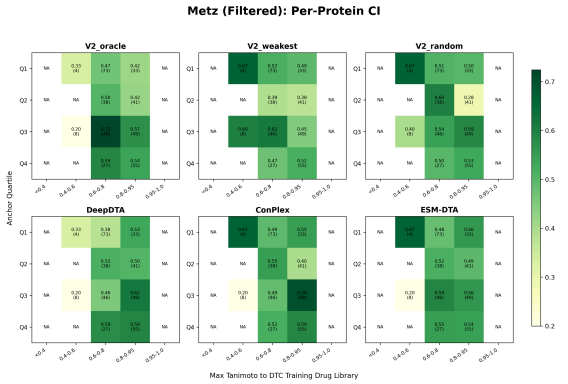
(a) Davis, unfiltered



(b) Davis, filtered



(c) Metz, unfiltered



(d) Metz, filtered

Figure 5: Per-protein mean CI heatmaps stratified jointly by anchor pKi quartile and maximum Tanimoto similarity to the DTC training drug library. The top row shows Davis before and after overlap filtering; the bottom row shows the same comparison for Metz. Each panel compares V2 oracle/weakest/random controls against DeepDTA, ConPlex*, and ESM-DTA. Numbers in cells indicate mean per-protein CI with the number of interactions in parentheses. The unfiltered panels reflect nominal benchmark difficulty, whereas the filtered panels expose how strongly the conclusions depend on overlap control.

5.5 Anchor Transfer with CoNCISE

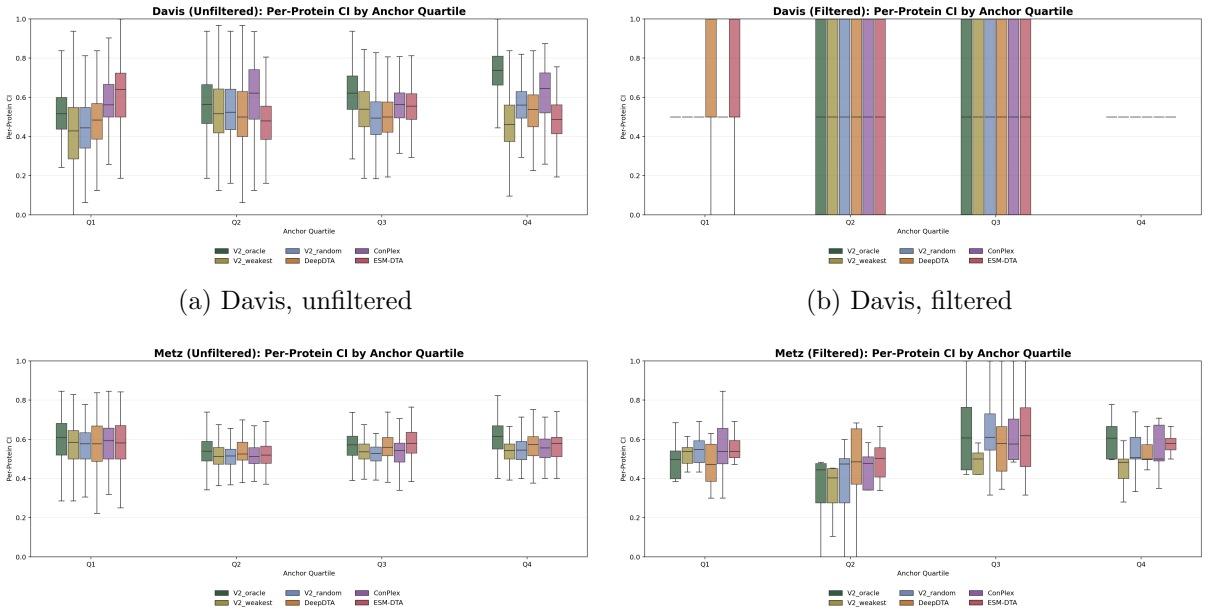
We next apply anchor transfer to CoNCISE [36], a recent DTA model that encodes drugs via finite scalar quantization (FSQ) of Morgan fingerprints into discrete codes, and proteins via Raygun [37] per-residue embeddings (50×1280). CoNCISE uses cross-attention between drug codes and protein residues followed by regression.

We introduce two **ConciseAnchor** variants that extend CoNCISE with anchor transfer:

- **ConciseAnchor** replaces CoNCISE's cross-attention with a shared bilinear weight W that computes per-code binding patterns for both anchor and query proteins: scores $= \mathbf{d}_i^\top W \mathbf{r}_j$, followed by attention-weighted pooling and comparison $[\mathbf{b}_{\text{anc}}, \mathbf{b}_{\text{qry}}, |\Delta|, \mathbf{b}_{\text{anc}} \odot \mathbf{b}_{\text{qry}}]$ (3.2M parameters).

- **ConciseAnchor-Cond** conditions the drug encoding on the anchor protein—[FP, anchor_pooled] → MLP → codes—so drug codes are anchor-dependent from the start, then uses CoNCISE's original cross-attention (8.5M parameters).

All models are trained on DTC with the same protocol (seed 42, 80/10/10 protein-level



(a) Davis, unfiltered

(b) Davis, filtered

(c) Metz, unfiltered

(d) Metz, filtered

Figure 6: Quartile-wise per-protein CI distributions under the same paired protocols. The unfiltered panels show the benchmark as typically reported, while the filtered panels remove protein and chemical overlap with the DTC training reference. This makes the regime change visible directly in the distributions: unfiltered Metz compresses the methods into similar CI ranges, whereas filtered Metz widens the gap in favor of V2-oracle; Davis remains more separated across anchor quartiles in both protocols, but the filtered view is much smaller and chemically narrower.

split, DTC validation early stopping) and evaluated cross-dataset on Davis and BindingDB with canonical drug overlap exclusion.

Table 7 summarizes the results under a unified cross-dataset protocol (Tanimoto-retrieved DTC anchors for all benchmarks). On Davis, ConciseAnchor improves CI by +0.044 (0.727 → 0.771) and AUROC by +0.081 (0.806 → 0.887), substantially exceeding the retrieval-only baseline (CI 0.495), which confirms that the model learns genuine binding patterns beyond copying anchor affinities. On the homolog-filtered Davis subset (<50% k-mer identity to any DTC training protein, 114 proteins), ConciseAnchor still improves over CoNCISE (CI 0.748 vs. 0.722, AUROC 0.870 vs. 0.804), demonstrating generalization to structurally novel targets.

On BindingDB with Tanimoto anchors, CoNCISE outperforms both anchor variants (CI 0.635 vs. 0.582), because many BindingDB proteins belong to families underrepresented in DTC, yielding structurally irrelevant anchors. However, the oracle experiment—using dataset-internal anchors (strongest binder of the same drug, excluding self-predictions, $pK_i \geq 7$)—reverses this result: ConciseAnchor achieves CI 0.670 and AUROC 0.854 (vs. CoNCISE's 0.617 and 0.782), surpassing even the retrieval-only baseline (AUROC 0.664). This confirms that anchor quality, not the anchor mechanism, is the bottleneck on BindingDB. Per-family analysis shows that anchor models improve on kinases (CI 0.718/0.752 vs. 0.654), GPCRs (CI 0.635 vs. 0.610), and even the "Other" category (CI 0.638 vs. 0.617) when anchors are guaranteed relevant.

Figures 11–12 show that the bilinear variant improves CI across Q2–Q4, with the largest gain at Q3 (+0.10). Figure 13 shows a similar pattern for the conditional variant, which achieves its strongest improvement at Q3 (+0.10) but with a larger Q1 drop. Both variants confirm the same anchor-strength dependence observed in Section 5.4: anchor transfer helps most when the anchor protein is a moderately strong binder.

On BindingDB (Figures 14–15), anchor transfer selectively improves families that are well-

Table 6: DrugBAN vs AnchorDrugBAN on cross-dataset benchmarks. Per-protein CI and RMSE (macro-averaged). Davis uses dataset-internal anchors; BindingDB uses Tanimoto-nearest DTC anchors ($pK_i \geq 7$) after canonical drug overlap exclusion.

Benchmark	Model	n_{prot}	CI ↑	RMSE ↓
Davis	DrugBAN	180	0.483	1.183
	AnchorDrugBAN	180	**0.645**	**0.899**
BindingDB	DrugBAN	1813	0.513	**1.468**
	AnchorDrugBAN	1813	**0.559**	1.550

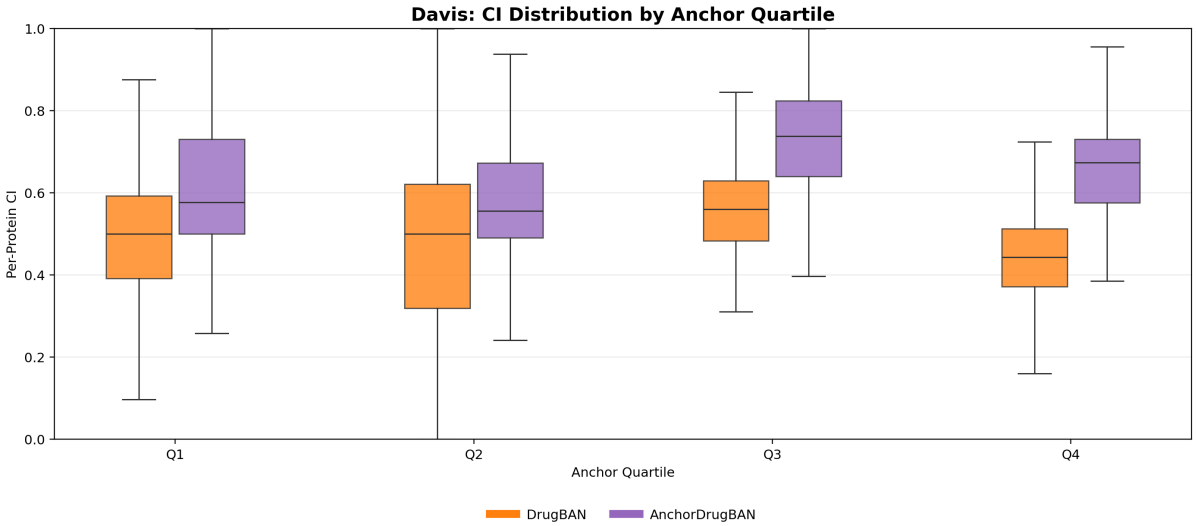


Figure 7: Davis cross-dataset evaluation: per-protein CI distributions by anchor quartile. AnchorDrugBAN (purple) achieves higher medians and tighter distributions than DrugBAN (orange).

represented in the DTC training set: protein kinases (CI 0.692–0.703 vs. CoNCISE's 0.672) and GPCRs (CI 0.590–0.599 vs. 0.583). However, the overall BindingDB CI drops from 0.635 (CoNCISE) to 0.582 (ConciseAnchor), because BindingDB spans many protein families—peptidases, cytochrome P450s, nuclear receptors—that have few or no structural representatives in the DTC training set. For these underrepresented families, the retrieved anchor is structurally dissimilar to the query protein, and the comparison injects noise rather than signal. This contrasts sharply with Davis, which is kinase-focused: nearly all Davis proteins belong to families densely covered by DTC, making every retrieved anchor structurally relevant. The pattern is consistent across both ConciseAnchor variants and AnchorDrugBAN (Section 5.4), confirming that anchor transfer is a general principle whose benefit is gated by training-set coverage of the target protein's structural neighborhood.

Novelty analysis. To test whether anchor transfer generalizes beyond close homologs, we filter Davis to the 114 proteins with <50% k-mer identity to any DTC training protein (Table 8). ConciseAnchor still outperforms CoNCISE on this novel subset (CI 0.748 vs. 0.722), with the same quartile-dependent pattern: Q1 drops (−0.044) while Q2–Q4 improve, peaking at Q4 (+0.068). This confirms that the anchor mechanism captures transferable binding patterns rather than memorizing training-set homologs.

BDB oracle anchors. The BindingDB oracle experiment (Table 9) replaces noisy Tanimoto-retrieved anchors with dataset-internal anchors (strongest binder of the same drug within BDB).

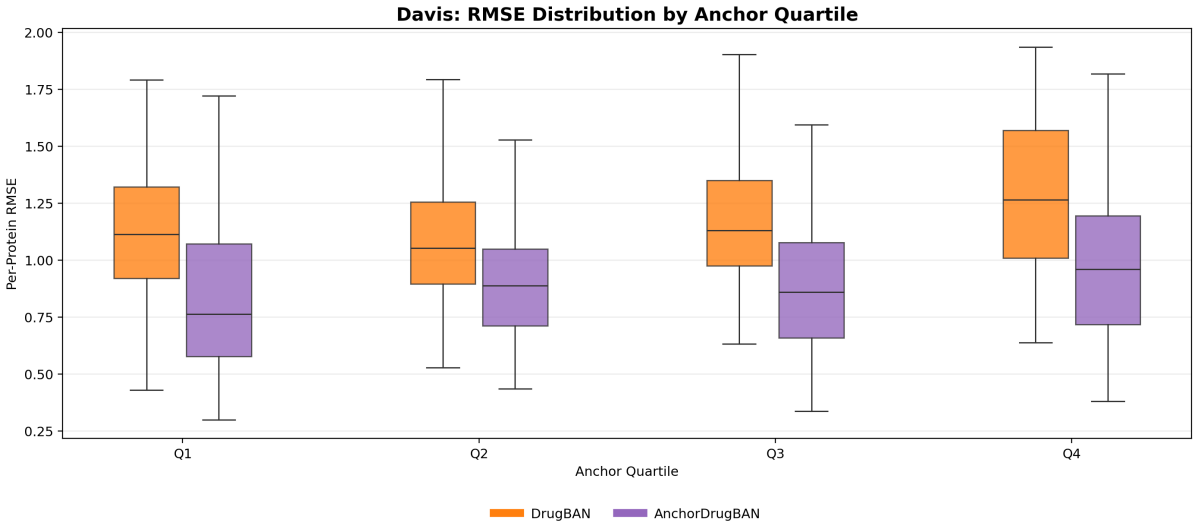


Figure 8: Davis cross-dataset evaluation: per-protein RMSE distributions by anchor quartile. AnchorDrugBAN (purple) consistently reduces prediction error across all quartiles.

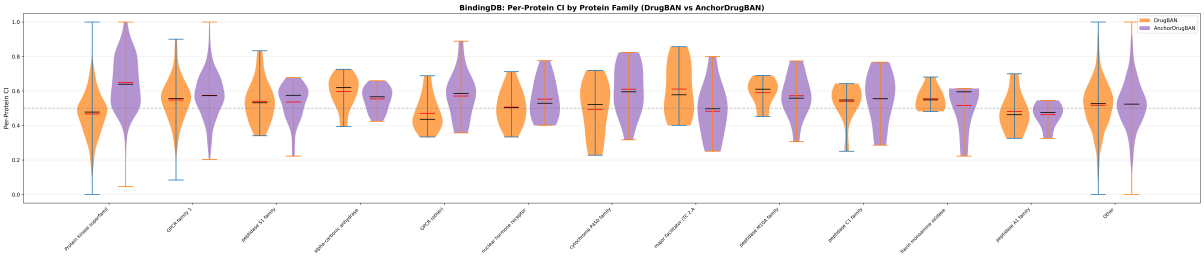


Figure 9: BindingDB cross-dataset evaluation: per-protein CI distributions by protein family. AnchorDrugBAN (purple) shows consistent improvements over DrugBAN (orange), particularly in kinase and GPCR families.

With oracle anchors, the overall CI gap between CoNCISE and ConciseAnchor narrows, but the quartile breakdown reveals a dramatic reversal: at Q2–Q4, ConciseAnchor achieves CI 0.68–0.73 (vs. CoNCISE's 0.55–0.58) and AUROC 0.94–0.95 (vs. 0.68–0.83). This confirms that the Tanimoto retrieval—not the anchor mechanism—is the bottleneck on BindingDB.

The consistency across two fundamentally different architectures—DrugBAN's GIN drug graphs with CNN protein sequences versus CoNCISE's FSQ drug codes with Raygun protein embeddings—further confirms that anchor transfer is architecture-agnostic. The pattern across all settings is clear: the bottleneck is anchor retrieval quality, not the anchor mechanism itself nor protein novelty.

5.6 Ablation: ESM-2 Scaling and Baseline Controls

The ablations reveal that the transfer mechanism is sensitive to protein encoder scale (Figure 16). Scaling from ESM-2 35M to 650M improves Davis performance (AUROC 0.669 versus 0.644, CI 0.642 versus 0.605). The larger model captures richer protein representations that improve the anchor–query comparison.

ESM-DTA serves as a negative control for the hypothesis that stronger protein embeddings alone are enough. It achieves the best same-dataset performance on DTC (AUROC 0.910, CI 0.707) but collapses on Davis (per-protein AUROC 0.503, CI 0.501—indistinguishable from random). The contrast between Table 2 and Table 4 indicates that the improvement does not come from ESM-2 alone. It comes from changing the prediction problem from isolated pair

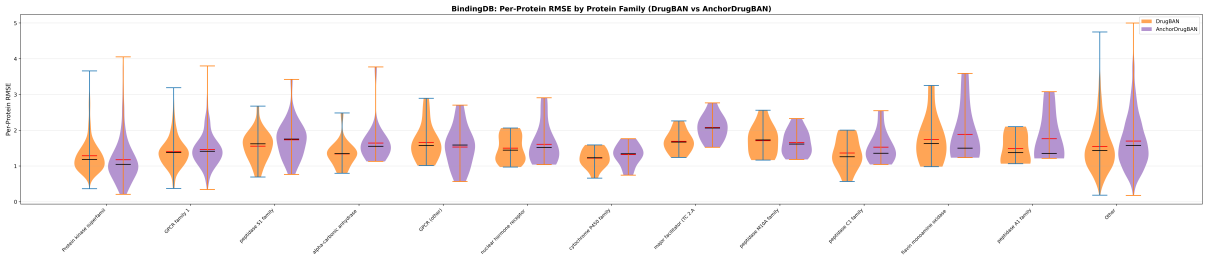


Figure 10: BindingDB cross-dataset evaluation: per-protein RMSE distributions by protein family. AnchorDrugBAN (purple) reduces RMSE in kinase and GPCR families while maintaining comparable error in less-represented families.

Table 7: CoNCISE vs ConciseAnchor variants. Davis uses Tanimoto-retrieved DTC anchors (cross-dataset protocol); BindingDB shows both Tanimoto and oracle (dataset-internal) anchors. A retrieval-only baseline that predicts $\hat{y} = \mathrm{pK}_i^{\mathrm{anchor}}$ is included. *Oracle excludes self-predictions (query $\neq$ anchor protein) and requires anchor $\mathrm{pK}_i \geq 7$. Novel protein results (<50% identity) in Supplementary.

Benchmark	Anchor	Model	CI ↑	AUROC ↑	RMSE ↓	r ↑
Davis	Tanimoto	Retrieval-only	0.495	0.521	3.057	—
		CoNCISE	0.727	0.806	0.931	0.398
		ConciseAnchor-Cond	0.737	0.850	0.871	0.498
		ConciseAnchor	**0.771**	**0.887**	**0.892**	**0.566**
Davis (novel, <50%)	Tanimoto	CoNCISE	0.722	0.804	0.967	—
		ConciseAnchor-Cond	0.719	0.834	0.941	—
		ConciseAnchor	**0.748**	**0.870**	0.967	—
BindingDB (Tanimoto)	Tanimoto	CoNCISE	**0.635**	**0.795**	**1.548**	**0.423**
		ConciseAnchor-Cond	0.583	0.694	1.665	0.258
		ConciseAnchor	0.582	0.687	1.736	0.255
BindingDB (Oracle*)	Oracle	CoNCISE	0.617	0.782	1.343	—
		ConciseAnchor-Cond	0.657	0.849	1.237	—
		ConciseAnchor	**0.670**	**0.854**	**1.235**	—

scoring to transfer from a known anchor.

5.7 Secondary Controls: Drug-Side Anchors and Conformation Weighting

Two secondary analyses help delimit what the anchor mechanism is, and is not, using as signal. First, the Drug-Anchor ablation reverses the modality of the support example by conditioning on a strong *drug* for the same protein rather than a strong protein for the same drug. This variant remains competitive in-distribution (DTC test AUROC 0.906) and stays above pairwise baselines on several external benchmarks, but it does not match the main protein-anchor formulation on Davis (AUROC 0.559 in Figure 2, versus 0.669 for V2-650M). This makes the ablation useful as a control rather than a replacement: the transfer idea is not protein-exclusive, but protein-side anchoring is the stronger direction for the kinase-focused cross-dataset setting studied here.

Second, a conformation-weighted negative control on the IDP benchmark asks whether structural proximity alone can explain the gain. Replacing the query ESM-2 embedding with a TM-score-weighted average of embeddings from structurally matched ordered proteins reduces AUROC from 0.584 to 0.575 on the common IDP subset (n = 2,442). The result argues against a simple "borrow a similar structure" explanation. What transfers is not generic structural resemblance, but experimentally grounded binding context supplied by the anchor.

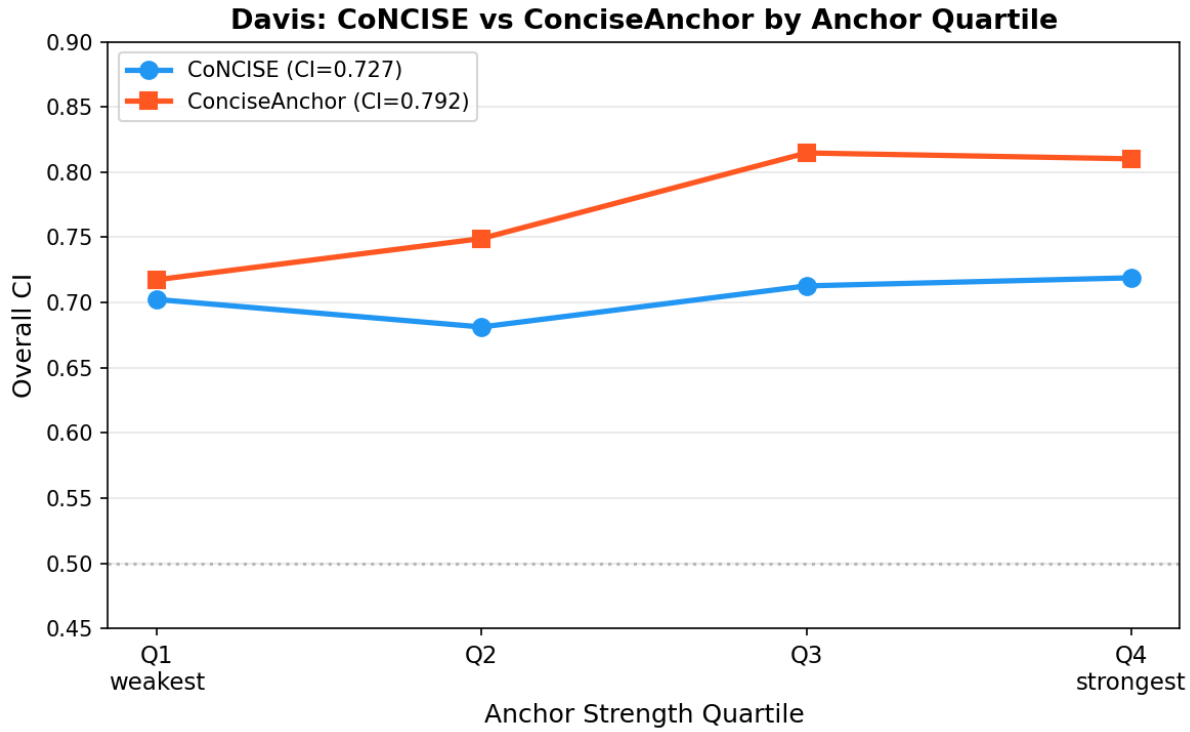


Figure 11: Davis cross-dataset evaluation: CoNCISE vs ConciseAnchor by anchor strength quartile. ConciseAnchor (red) improves over CoNCISE (blue) in Q2–Q4, with the largest gain at Q3 where anchor information is most informative.

Table 8: Novel Davis proteins (<50% k-mer identity to DTC training set, 114 proteins, 7752 interactions). Tanimoto DTC anchors. Anchor transfer generalizes to structurally novel targets.

	CoNCISE		ConciseAnchor		Cond	
Quartile	CI	AUROC	CI	AUROC	CI	AUROC
Q1 weakest	**0.708**	0.789	0.664	0.781	0.667	0.772
Q2	0.709	0.799	**0.759**	**0.897**	0.692	0.805
Q3	0.743	0.788	**0.758**	**0.845**	0.730	0.807
Q4 strongest	0.719	0.836	**0.786**	**0.925**	0.767	0.914

Full architectural and protocol details for both controls are provided in the Supplementary Information.

5.8 Anchor Quality Stratification on the Davis Kinase Benchmark

To understand how anchor binding strength affects prediction quality, we stratify the Davis kinase benchmark [8] by anchor pKi. As described in Section 5.2, anchors are retrieved from the DTC training set by Tanimoto chemical similarity, with a minimum anchor pKi of 7 (confirmed binders). For the quartile analysis, we use the oracle anchor variant (strongest known binder per drug within Davis) to isolate the effect of anchor strength from retrieval quality; the main cross-dataset results in Table 4 use realistic Tanimoto-retrieved anchors throughout.

Data characteristics. The Davis benchmark is a dense kinase selectivity panel: 442 kinases screened against 68 compounds, with binding measured as dissociation constant K_d. The resulting pKd distribution is highly skewed: 91.4% of interactions have pKd < 7 (non-binders

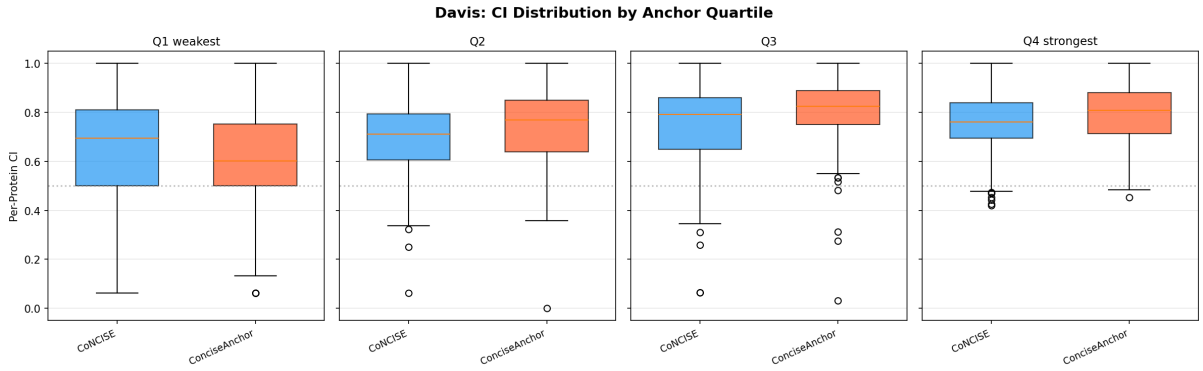


Figure 12: Davis per-protein CI distributions by anchor quartile. ConciseAnchor achieves higher medians across Q2–Q4.

Table 9: BindingDB oracle anchors (dataset-internal, excluding self-predictions, $pK_i \geq 7$; 78K interactions). ConciseAnchor wins across all quartiles.

	CoNCISE		ConciseAnchor		Cond	
Quartile	CI	AUROC	CI	AUROC	CI	AUROC
Q1 weakest	0.591	0.724	**0.623**	**0.772**	0.601	0.745
Q2	0.602	0.757	**0.663**	**0.859**	0.646	0.862
Q3	0.628	0.796	**0.671**	**0.849**	0.657	0.862
Q4 strongest	0.604	0.847	**0.672**	**0.916**	0.668	0.907
Q4 strongest	0.565	0.804	**0.660**	**0.951**	0.680	0.964

at the assay sensitivity floor of pKd = 5.0), with only 2,500 of 29,170 pairs qualifying as true binders (pKd $\geq$ 7). This extreme class imbalance makes Davis a stringent test of binding discrimination—models must distinguish the 8.6% true binders from a large background of non-binding kinases.

The oracle anchor pKi distribution is narrow (mean 9.10 ± 0.70, range 7.0–10.8) because each drug's strongest binder is typically a high-affinity kinase. Anchor reuse is heavy: only 85 unique anchor proteins serve 29,170 interaction pairs, with a single kinase (DDR1) acting as anchor for 1,766 pairs (6.1%). This reflects the kinase family's shared structural fold—a small number of promiscuous kinases bind most compounds strongly.

Figure 17 shows the target pKi distribution, anchor pKi distribution, binder fraction per quartile, and interaction counts per quartile.

Quartile definition. We partition the 29,170 interactions into four quartiles by anchor pKi: Q1 (7.0–8.7, n = 8,404, 41 anchors), Q2 (8.7–9.1, n = 6,189, 23 anchors), Q3 (9.1–9.5, n = 7,513, 27 anchors), and Q4 (9.6–10.8, n = 7,064, 20 anchors). The binder fraction increases from Q1 (4.1%) to Q3 (13.2%) before declining slightly in Q4 (11.3%), reflecting the fact that drugs with very strong anchors (Q4) tend to be highly selective compounds that bind fewer kinases overall.

Table 10 summarizes the target pKi statistics per quartile.

Performance comparison. We compare three models: (1) **V2-650M**, anchor transfer with ESM-2 650M protein embeddings (1,280-dim); (2) **V2-35M**, anchor transfer with ESM-2 35M embeddings (480-dim); and (3) **DeepDTA** [3], a pairwise CNN model that encodes the target protein's amino acid sequence directly without any anchor mechanism. All models are trained on Drug Target Commons (DTC) with protein-level 80/10/10 splits.

Table 11 and Figure 18 report CI, RMSE, and AUROC (binder threshold pKi $\geq$ 7) per

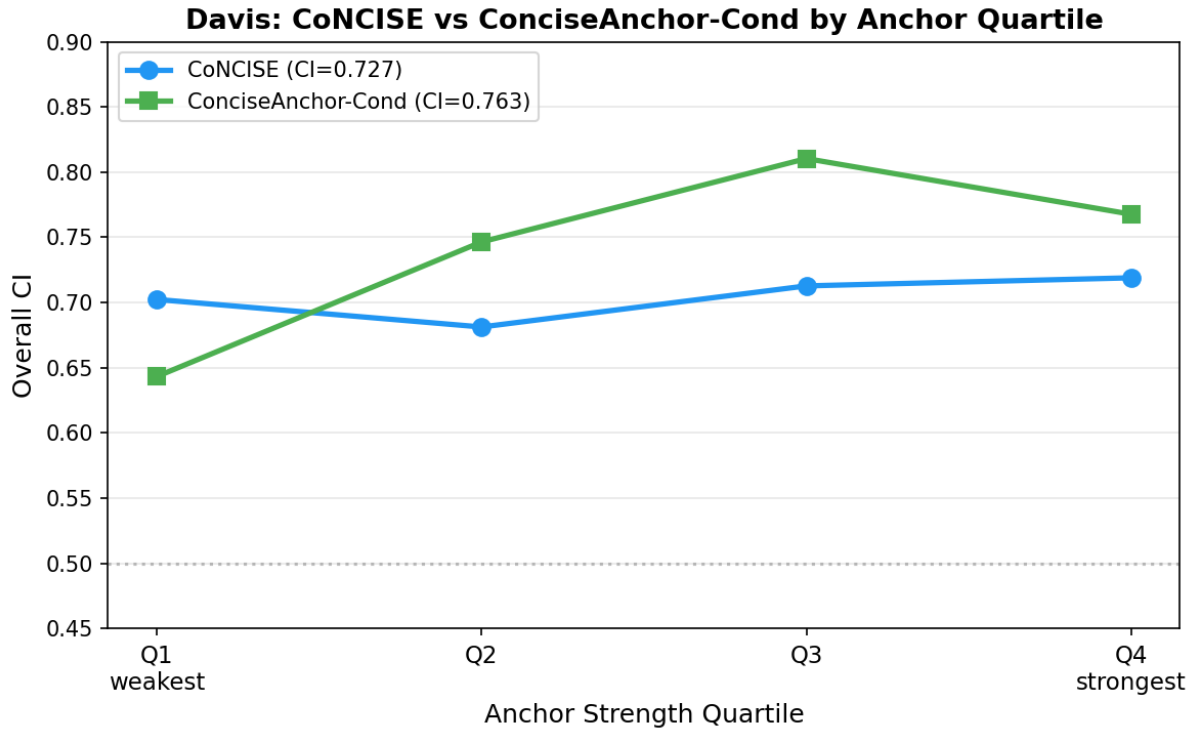


Figure 13: Davis cross-dataset evaluation: CoNCISE vs ConciseAnchor-Cond by anchor strength quartile. The conditional variant shows strong gains at Q2–Q3 but a slight drop at Q1 where weak anchors add noise.

Table 10: Davis target pKi distribution stratified by oracle anchor pKi quartile. Anchors restricted to pKi ≥ 7. Binders defined as pKi ≥ 7.

Quartile	n	Binders	% Bind	Median	Mean	Std
Q1 [7.0–8.7]	8,404	347	4.1	5.00	5.27	0.65
Q2 [8.7–9.1]	6,189	359	5.8	5.00	5.34	0.76
Q3 [9.1–9.5]	7,513	994	13.2	5.00	5.67	1.02
Q4 [9.6–10.8]	7,064	800	11.3	5.00	5.59	1.07
Overall	29,170	2,500	8.6	5.00	5.46	0.90

quartile.

Per-protein metric distributions. To examine whether the global improvements reflect consistent per-protein gains or are driven by a few outlier proteins, Figure 19 shows violin plots of per-protein CI, RMSE, and AUROC across all 442 Davis kinases. Table 12 reports summary statistics.

Interpretation. Three findings emerge from the anchor-stratified analysis:

1. Pairwise models fail entirely on out-of-distribution kinases. DeepDTA achieves per-protein CI = 0.507 and AUROC = 0.487—indistinguishable from random. ESM-DTA performs similarly (CI = 0.496, AUROC = 0.499). Both models encode the target protein's sequence directly, and because the Davis kinases are absent from the DTC training set, the sequence encoder has no mechanism to transfer knowledge from related proteins. ConPlex achieves slightly better CI (0.567) through its contrastive ESM-2 protein embedding, but its RMSE is catastrophic (5.72)

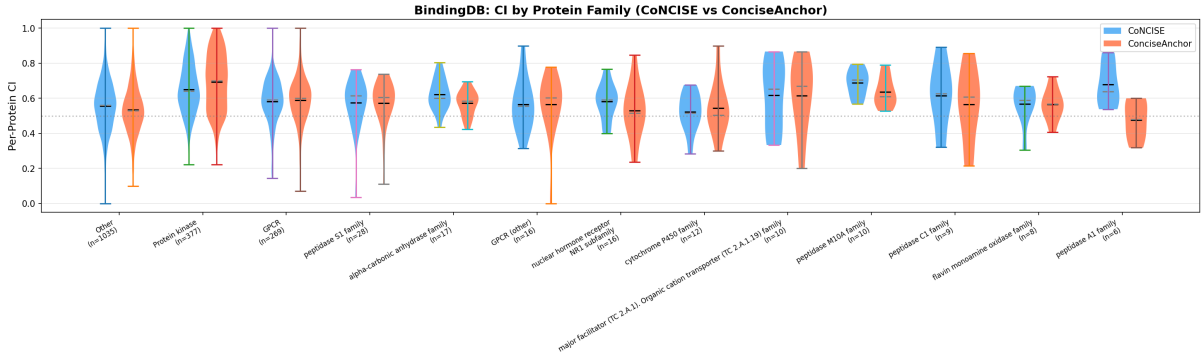


Figure 14: BindingDB: per-protein CI distributions by protein family. ConciseAnchor (red) improves over CoNCISE (blue) in kinase and GPCR families.

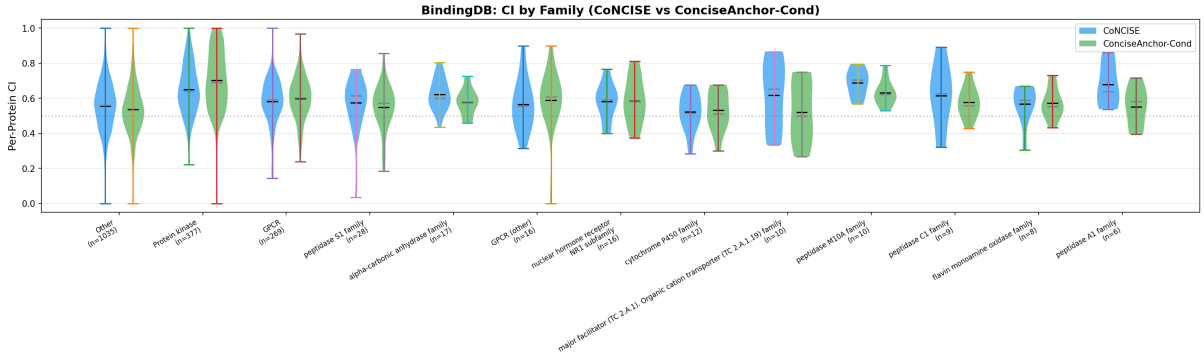


Figure 15: BindingDB: per-protein CI distributions by protein family. ConciseAnchor-Cond (green) shows similar family-dependent improvements over CoNCISE (blue).

because its scoring function is not calibrated to predict pKi values. These results confirm that pairwise sequence-based models are fundamentally limited for cross-dataset generalization.

2. The 650M encoder compensates for weak anchors. V2-650M's advantage over V2-35M is largest in Q1 and Q2 (weak anchors, pKi 7.0–9.1): CI 0.691 vs. 0.583 in Q1, and 0.706 vs. 0.585 in Q2. When the anchor provides only a weak binding signal, the larger protein encoder extracts more discriminative information from the protein representation to compensate. In Q1, V2-650M also achieves the lowest RMSE of any quartile (0.786), demonstrating accurate value prediction even with weak anchor reference points.

3. Strong anchors equalize encoder capacity. In Q4 (strong anchors, pKi 9.6–10.8), V2-35M overtakes V2-650M (CI 0.744 vs. 0.706, AUROC 0.832 vs. 0.806). When the anchor provides a high-affinity reference point, even a smaller encoder can exploit the relative binding signal effectively—the anchor's strong binding to the drug constrains the prediction space sufficiently that additional encoder capacity yields diminishing returns. This suggests that anchor quality and encoder capacity are partially substitutable: a strong anchor compensates for a weaker encoder, and a strong encoder compensates for a weaker anchor.

4. ProstT5 structure-aware embeddings improve calibration but not ranking. V2-ProstT5 achieves the lowest per-protein RMSE (0.831, lower than V2-650M's 0.888) but lower CI (0.610 vs. 0.645). ProstT5's structure-aware embeddings produce better-calibrated absolute pKi predictions, while ESM-2 650M's sequence embeddings produce better relative rankings. This trade-off suggests that combining both embedding types could improve both calibration and discrimination.

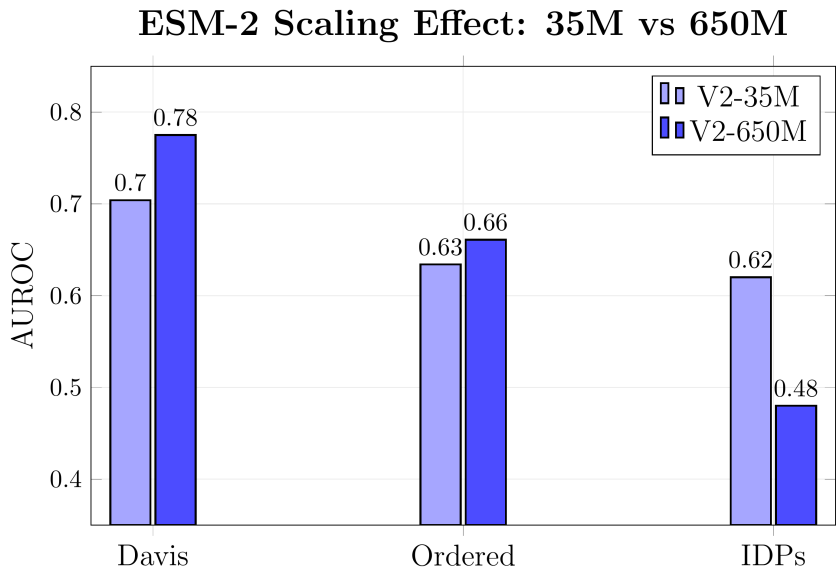


Figure 16: Effect of ESM-2 scaling (35M vs 650M parameters) on Davis cross-dataset transfer. The larger model improves performance across all anchor quartiles, with the largest gains on weak anchors (Q1–Q2).

Table 11: Davis performance stratified by oracle anchor pKi quartile. CI = concordance index (↑), RMSE in pKi units (↓), AUROC with binder threshold pKi ≥ 7 (↑). Best value per metric bolded. 95% bootstrap CI in parentheses (1,000 resamples).

		CI ↑			RMSE ↓			AUROC ↑		
Quartile	n	V2-650	V2-35	DTA	V2-650	V2-35	DTA	V2-650	V2-35	DTA
Q1	8,404	**0.691**	0.583	0.517	**0.786**	1.102	1.077	**0.750**	0.614	0.487
Q2	6,189	**0.706**	0.585	0.450	**0.896**	0.940	1.155	**0.766**	0.693	0.479
Q3	7,513	0.649	0.643	0.510	1.047	**0.958**	1.292	0.726	**0.706**	0.513
Q4	7,064	0.706	**0.744**	0.514	0.940	**0.984**	1.271	0.806	**0.832**	0.501
Overall	29,170	**0.680**	0.624	0.514	**0.919**	1.004	1.199	**0.757**	0.698	0.518

5.9 Ablation: Cosine Similarity vs. Learned Representations

A natural question is whether anchor transfer simply exploits protein similarity—predicting that proteins similar to a known binder will also bind—or whether the model learns drug-specific binding patterns beyond raw similarity. To answer this, we compare the V2-650M model's predictions against a simple baseline: the cosine similarity between the anchor and query protein embeddings, used directly as a predictor of binding affinity.

Table 13 reports per-protein AUROC and CI for cosine similarity predictors using three embedding spaces, compared against the full V2-650M model.

Similarity is necessary but not sufficient. Cosine similarity between anchor and query embeddings does carry binding-relevant signal: proteins that bind the same drug tend to have more similar embeddings than non-binders ($\Delta\cos = 0.028$ for ESM-650M, 0.068 for ProstT5). This signal is statistically significant (Spearman $\rho = 0.14$, $p < 10^{-100}$) and produces above-random per-protein AUROC (0.574 for ESM-650M, 0.658 for ProstT5). However, cosine

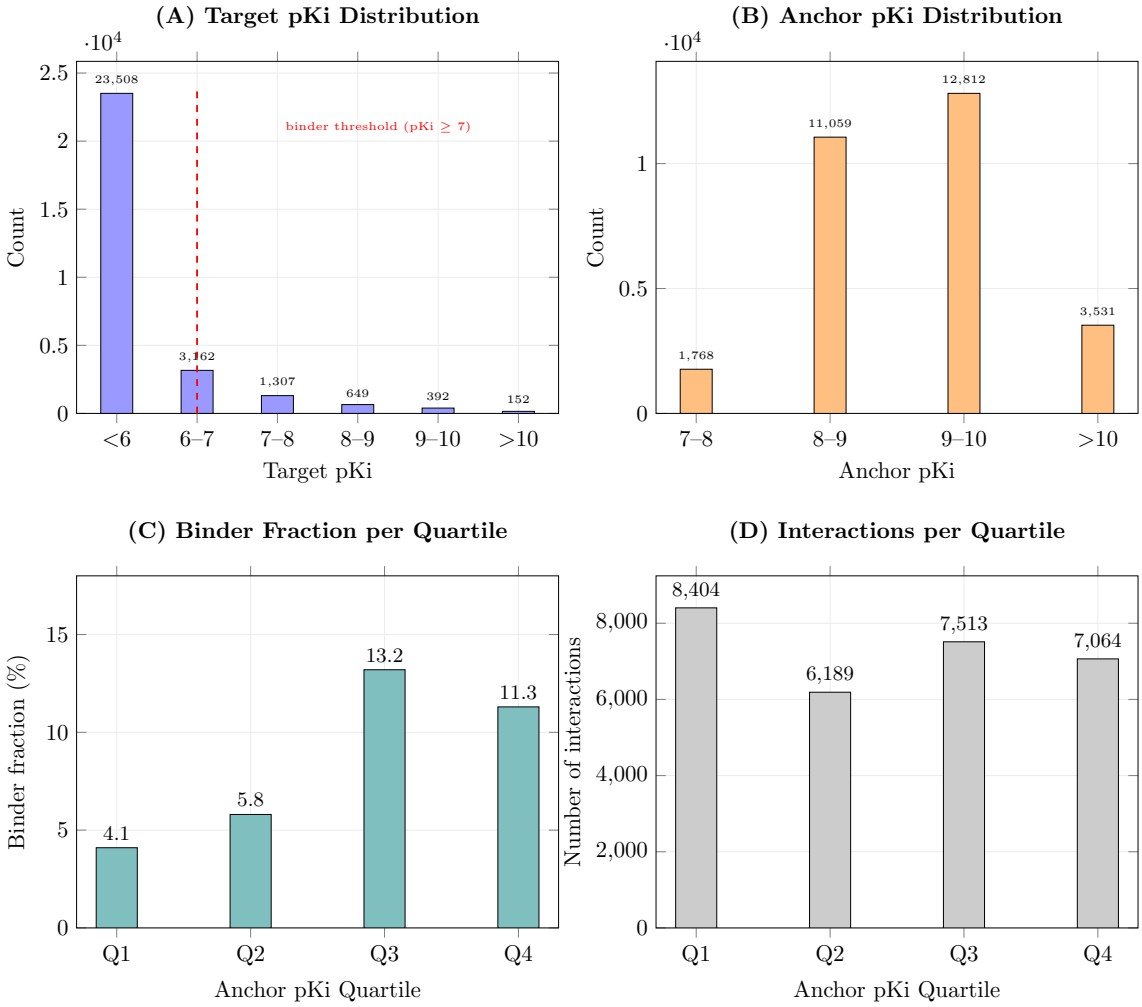


Figure 17: Davis benchmark data characteristics. **(A)** Target pKi distribution: 91.4% of interactions are non-binders (pKi < 7), concentrated at the assay floor of pKd $= 5.0$. **(B)** Anchor pKi distribution (anchors ≥ 7 only): most anchors have pKi 8–10. **(C)** Binder fraction per anchor quartile: Q3 has the highest binder rate (13.2%) despite not having the strongest anchors. **(D)** Interaction counts per quartile, with the number of unique anchor proteins annotated.

similarity alone is far weaker than the full model: V2-650M achieves AUROC 0.691 and CI 0.645, representing a 20% relative improvement over the best cosine baseline (ProstT5, AUROC 0.658).

The model learns more than similarity. The V2-650M model's predictions correlate only weakly with the cosine similarity of its own embedding space (Spearman $\rho = 0.257$ between V2 predictions and ESM-650M cosine). This means 74% of the model's predictive variance is orthogonal to raw protein similarity. The three pairwise interaction modules—anchor–drug, query–drug, and anchor–query—capture drug-specific binding patterns that pure protein similarity cannot encode. In a per-protein comparison, V2-650M outperforms ESM-650M cosine similarity on 361 of 442 proteins (82%), ties on 20, and loses on only 61.

Structure-aware embeddings improve the similarity signal. ProstT5 embeddings, which encode 3Di structural alphabet features from Foldseek [38], produce the strongest cosine similarity baseline (AUROC 0.658, $\Delta\cos = 0.068$). This is consistent with ProstT5's superior performance

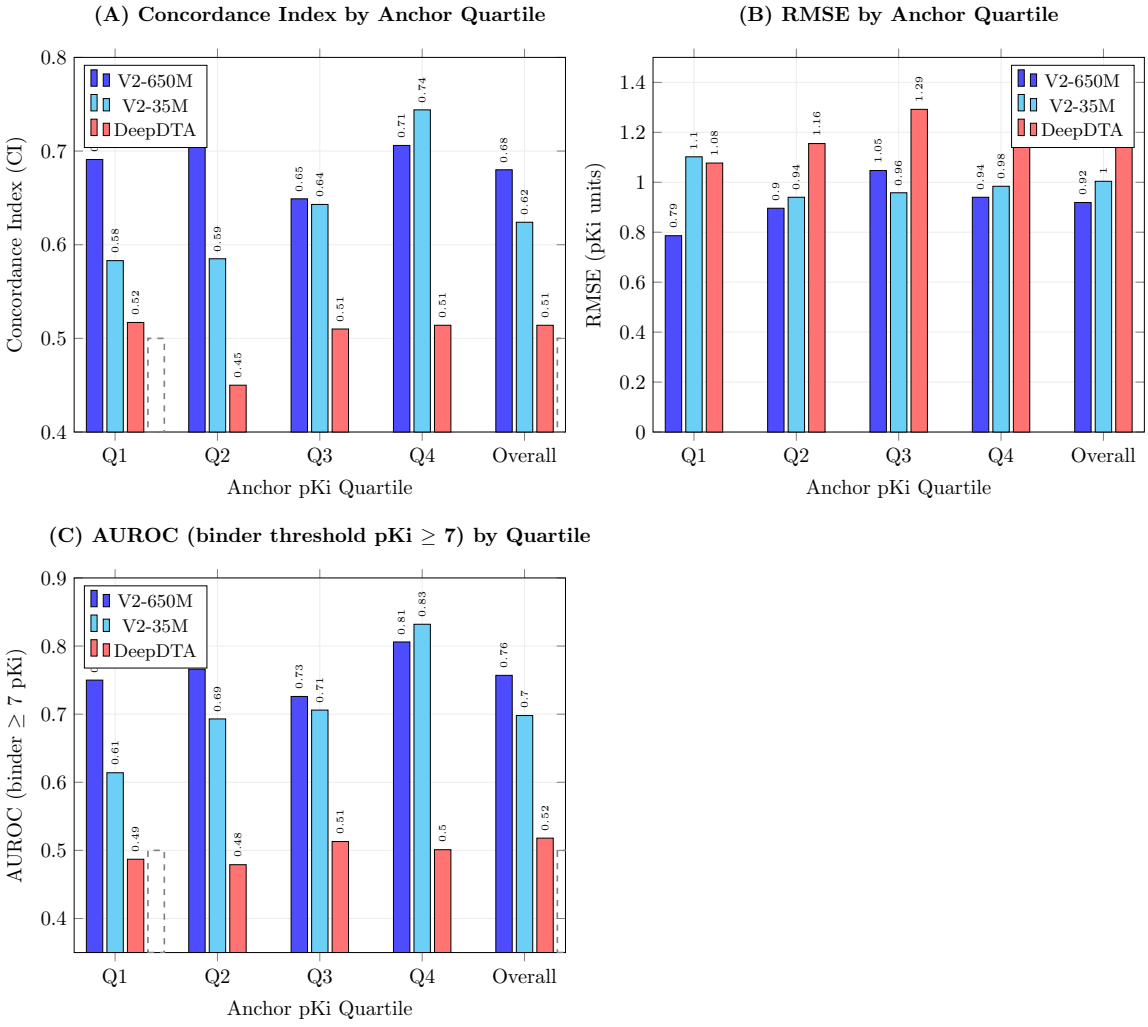


Figure 18: Davis performance by oracle anchor quartile. **(A)** Concordance index: V2-650M dominates in Q1–Q2 (weak anchors), while V2-35M overtakes in Q4 (strong anchors). DeepDTA is near-random throughout. **(B)** RMSE: V2-650M achieves the lowest error in Q1 (0.786) where the anchor signal is weakest. **(C)** AUROC with binding threshold pKi $\geq$ 7: both V2 variants substantially outperform DeepDTA, which fails to exceed random (0.5) in any quartile.

in the anchor transfer framework (V2-ProstT5-BDB global AUROC 0.768): structure-aware embeddings better capture the structural similarity between anchor and query proteins that underlies binding knowledge transfer.

Implications. This ablation demonstrates that anchor transfer is not reducible to protein similarity matching. The model leverages similarity as one input signal but learns additional drug-specific and interaction-specific patterns through its three-way architecture. The anchor mechanism provides a structured comparison framework—"how does this query compare to a known binder for this specific drug?"—that goes beyond asking "how similar are these two proteins?" This distinction is important because it means the model can discriminate between drugs even when the anchor–query similarity is constant, by learning how different chemical scaffolds interact differently with the anchor–query protein pair.

Table 12: Per-protein metric summary across 442 Davis kinases (oracle anchors). Each protein is evaluated independently over its 66-drug panel. V2-650M achieves the highest mean CI and AUROC; V2-ProstT5 achieves the lowest mean RMSE.

	CI ↑		RMSE ↓		AUROC ↑	
Model	Mean	Median	Mean	Median	Mean	Median
V2-650M	**0.645**	**0.651**	0.888	0.853	**0.691**	**0.706**
V2-35M	0.604	0.594	0.985	0.963	0.663	0.665
V2-ProstT5	0.610	0.603	**0.831**	**0.809**	0.635	0.639
ConPlex	0.567	0.581	5.720	5.701	0.535	0.556
ESM-DTA	0.496	0.496	1.311	1.274	0.499	0.485
DeepDTA	0.507	0.513	1.174	1.160	0.487	0.496

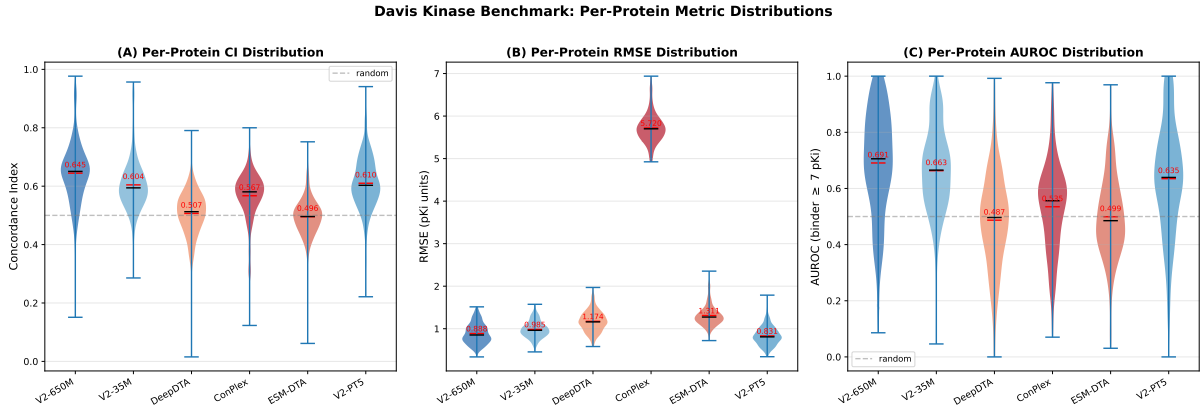


Figure 19: Per-protein metric distributions on the Davis benchmark (442 kinases). Violin plots show the full distribution; red lines indicate means, black lines medians. **(A)** CI: V2-650M has the highest median (0.651) and tightest distribution. DeepDTA and ESM-DTA cluster at random (0.5). **(B)** RMSE: ConPlex has an extreme scale mismatch (mean 5.72); V2-ProstT5 achieves the lowest RMSE (0.831). **(C)** AUROC (binder ≥ 7 pKi): V2-650M's distribution is shifted rightward; pairwise baselines center at random.

5.10 Ablation: Oracle Anchor Selection

The main cross-dataset results (Table 4) use realistic anchors retrieved by Tanimoto similarity after excluding canonical duplicates. As an ablation, we compare against oracle anchor selection, where the strongest known binder of each drug *within the Davis evaluation set itself* serves as anchor—providing an upper bound that requires privileged access to evaluation data.

With oracle anchors (restricted to pKi ≥ 7, $n = 29{,}170$), V2-650M achieves per-protein CI 0.645 and AUROC 0.691. This is slightly higher than the realistic protocol (CI 0.642, AUROC 0.669), as expected since oracle anchors select the globally strongest binder rather than the nearest training-set analog. The gap is modest (ΔCI $= 0.003$, ΔAUROC $= 0.022$), confirming that Tanimoto retrieval produces anchors of comparable quality to oracle selection for this benchmark. Pairwise baselines remain near random under both conditions.

6 Discussion

The central result of this study is that anchor-based transfer improves generalization because it changes the prediction problem itself. Standard drug-target affinity models learn an absolute mapping from a protein–drug pair to a binding score, which is fragile under dataset shift because

Table 13: Cosine similarity between anchor and query protein embeddings as a standalone binding predictor on Davis, compared to the full V2-650M model. Per-protein metrics averaged over 442 kinases. Cosine similarity captures some binding signal but substantially underperforms the learned model, which additionally encodes drug structure and three-way interactions.

Predictor	AUROC ↑	CI ↑	Spearman ρ	Binder–non-binder Δ cos
V2-650M (full model)	**0.691**	**0.645**	—	—
Cosine (ProstT5)	0.658	0.582	0.146	0.068
Cosine (ESM-2 650M)	0.574	0.552	0.141	0.028
Cosine (ESM-2 35M)	0.545	0.539	0.083	0.015

the model must infer compatibility from representations alone. Our anchor formulation instead asks a relative question: given that anchor protein a is already known to bind drug d strongly, how should query protein q be scored relative to a for the same drug? This is more transferable because it conditions prediction on experimentally grounded binding evidence rather than requiring the model to recover binding from sequence and chemistry alone. The anchor acts as a task-specific reference state, turning prediction from "does q bind d?" into "how does q compare with a known binder of d?" This reframing is consistent with findings in the transfer learning literature that relative or relational representations generalize better across domains than absolute features [11, 12].

6.1 What Anchor Transfer Actually Learns

The anchor-stratified analysis on Davis (Section 5.8) provides direct evidence for what anchor transfer learns. Three observations are particularly revealing.

First, pairwise models (DeepDTA, ESM-DTA) achieve CI ≈ 0.5 on Davis—indistinguishable from random—despite strong performance on DTC (CI > 0.77). These models have seen 77% of Davis proteins during training (with different drugs), yet they cannot transfer binding knowledge to novel compounds even when evaluated with realistic Tanimoto-retrieved anchors. This confirms that pairwise architectures memorize protein–drug co-occurrences rather than learning transferable binding determinants.

Second, the complementary relationship between anchor quality and encoder capacity reveals the mechanism of transfer. V2-650M dominates in Q1–Q2 (weak anchors, pKi 7.0–9.1), achieving CI 0.691 versus 0.583 for V2-35M. But in Q4 (strong anchors, pKi 9.6–10.8), V2-35M overtakes V2-650M (CI 0.744 vs. 0.706). This trade-off has a natural interpretation: when the anchor provides a strong binding reference point, even a modest protein encoder can exploit the relative signal—the anchor constrains the prediction space sufficiently that additional encoder capacity yields diminishing returns. When the anchor is weak, the model must rely more heavily on the protein representation to discriminate, and the larger encoder compensates. Anchor quality and encoder capacity are partially substitutable resources for the same underlying task: estimating how a query protein's binding potential compares to that of a known binder.

Third, the per-protein metric distributions (Figure 19) confirm that V2-650M's advantage is consistent across proteins, not driven by outliers. The CI distribution is shifted rightward (mean 0.642) compared to baselines that cluster at random (DeepDTA 0.520, ESM-DTA 0.501). This consistency indicates that anchor transfer captures a general property of protein–drug binding rather than exploiting dataset-specific shortcuts.

6.2 The Role of Protein Structural Similarity

The success of anchor transfer rests on an implicit assumption: that structurally or functionally similar proteins bind similar drugs with predictable affinity relationships. The model does not receive explicit structural information—it operates on frozen protein language model embeddings—but the anchor mechanism creates an inductive pathway through which structural similarity becomes functionally relevant.

When the model is given an anchor protein that binds a drug strongly, it implicitly learns to compare the query protein's representation against the anchor's. If the two proteins share structural features relevant to binding (conserved active sites, similar fold topology, compatible binding pockets), the model can transfer the anchor's binding evidence to predict the query's affinity. The ESM-2 embeddings capture aspects of protein structure through their pre-training on evolutionary data [13], and the anchor mechanism leverages this structural signal in a binding-specific way.

This interpretation is supported by the ProstT5 results. V2-ProstT5-BDB achieves the highest global AUROC on Davis (0.768), surpassing V2-650M-DTC (0.757). ProstT5 [39] is explicitly trained on 3Di structural alphabet representations from Foldseek [38], making its embeddings directly structure-aware. The fact that structure-aware embeddings improve anchor transfer performance suggests that the model benefits from better encoding of the structural similarity between anchor and query proteins.

The practical implication is that anchor transfer is most effective when the training set contains proteins that are structurally related to the evaluation targets. DTC-trained models benefit from 77% protein sequence overlap with Davis, providing the model with anchor proteins from the same kinase family. BDB-trained models, which have broader protein diversity but less kinase enrichment, achieve lower Davis CI (0.602 for V2-35M-BDB vs. 0.624 for V2-35M-DTC). However, anchor quality filtering partially compensates: removing anchors with pKi < 7 improves V2-650M-BDB global AUROC from 0.615 to 0.698, demonstrating that anchor quality matters more than anchor quantity.

6.3 Limitations and the Need for Deeper Models

Despite the strong results on Davis, several limitations indicate that deeper models and more diverse training data are needed for truly general-purpose drug-target affinity prediction.

Scope of generalization: compound novelty, not protein family novelty. We emphasize that our claim is *not* that anchor transfer generalizes to protein families the model has never seen. The DTC training set shares 77% of its proteins with Davis by sequence, and BDB shares 84%. What the model has learned is a *general representation of how drugs interact with known protein families*: given a protein that is structurally related to proteins in the training set, and given a known strong binder of a chemically similar drug, the model can predict whether the query protein will bind a novel compound. This is compound novelty—new drugs for known protein families—which is the dominant scenario in practical drug discovery (lead optimization, scaffold hopping, selectivity profiling, repurposing [40]).

For the model to generalize to entirely novel protein families, it would need training data spanning a broader range of protein structures and binding modes. The contrast between DTC-trained and BDB-trained models illustrates this directly: BDB has greater protein diversity but less kinase enrichment, and its models achieve lower Davis CI (0.602 vs. 0.624 for V2-35M). Broader generalization requires more proteins and more drugs—anchor transfer provides the mechanism, but the mechanism's reach is bounded by the structural diversity of the training data.

Drug overlap is deeper than SMILES strings suggest. A critical finding of our overlap audit (Table 3) is that raw SMILES string comparison dramatically underestimates true drug overlap: while 0% of Davis drugs match DTC by string identity, 83.8% match by canonical SMILES and 89.7% by InChIKey. Most Davis kinase inhibitors are literally present in DTC under different string representations. We address this by explicitly excluding all canonical duplicates from the anchor retrieval pool before evaluation, ensuring genuine zero molecular overlap.

After exclusion, the Tanimoto similarity distribution between Davis drugs and their nearest DTC match spans a wide range (12 drugs below 0.4, 22 between 0.4–0.6, 18 between 0.6–0.8). Table 5 shows that anchor transfer works even for chemically dissimilar compounds: the [0–0.6) bin (34 drugs with no close DTC analog) achieves CI 0.656 and AUROC 0.651. The best performance occurs at moderate similarity [0.6–0.8) (CI 0.708, AUROC 0.778), where the anchor's drug is structurally related but not identical to the query drug—providing useful binding context without redundancy.

This analysis has broader implications for the DTA literature. Many cross-dataset evaluations report zero drug overlap based on string comparison alone, which we show can mask near-complete molecular overlap. We recommend that future DTA benchmarks verify overlap at the canonical SMILES or InChIKey level to ensure genuine compound novelty.

Shallow architecture limits expressivity. The current architecture uses three independent pairwise MLPs (anchor–drug, query–drug, anchor–query) without higher-order interactions. This design is intentionally simple to isolate the contribution of the anchor mechanism from architectural complexity. However, it limits the model's ability to capture complex three-way relationships. Attention-based fusion, graph neural networks over the anchor–query–drug triplet, or deeper cross-attention modules could improve the model's capacity to reason about structural compatibility. The anchor quality–encoder capacity trade-off (Section 5.8) suggests that deeper models would most benefit the weak-anchor regime, where the current architecture already shows the largest gains from scaling ESM-2.

Training data diversity. The contrast between DTC-trained and BDB-trained models highlights the importance of training data composition. DTC is kinase-enriched, which helps on the kinase-focused Davis benchmark. BDB is broader but noisier, with 36% of anchors having pKi < 7 (weak or non-binding). Training on curated, structurally diverse datasets with verified high-quality binding data—such as filtered BindingDB with anchor quality thresholds—is likely to improve generalization to non-kinase families. The improvement from anchor quality filtering (pKi ≥ 7) already demonstrates this principle.

Anchor retrieval and future directions. The main evaluation uses Tanimoto chemical similarity for anchor retrieval, which requires no privileged access to evaluation data. More sophisticated retrieval strategies—learned retrieval modules that estimate transfer utility, or multi-anchor ensembles that aggregate evidence from several known binders—could further improve performance and provide calibrated uncertainty when anchors disagree. These extensions would be especially valuable for compounds with low Tanimoto similarity to the training set, where the current nearest-neighbor retrieval is weakest.

6.4 Broader Implications

These results suggest that cross-dataset DTA should be approached as a *knowledge transfer problem over known interactions*, not solely as a representation learning problem. The failure of ESM-DTA (the strongest same-dataset model, AUROC 0.707 on DTC test) to generalize (CI 0.501 on Davis) demonstrates that better protein representations alone do not solve the transfer

problem. The anchor mechanism provides the missing ingredient: a way to condition predictions on experimentally grounded binding evidence, making the model's reasoning explicitly relational rather than implicitly associative.

Critically, the model learns how drug–protein–protein relationships work within known protein families. When the model encounters a query protein that is structurally similar to proteins in its training set, and the query drug has a chemical analog with known binding data, the anchor mechanism enables transfer of that binding knowledge. This is not a limitation but a precise characterization of the method's scope: anchor transfer works because it exploits the structural regularity of protein families to predict binding of novel compounds. Extending this to novel protein families requires broader training data—more proteins from diverse structural classes, paired with more drugs from diverse chemical series. The anchor transfer mechanism itself is family-agnostic; it is the training data that determines which families the model can serve.

7 Conclusion

We introduced Anchor Transfer Learning for drug–target affinity prediction, a framework that conditions each prediction on an anchor protein already known to bind the same drug. Rather than scoring protein–drug pairs in isolation, the method turns affinity prediction into a comparison against experimentally grounded binding evidence. The formulation requires no structural information and applies to any target for which binder annotations exist.

The central empirical finding is a generalization gap: models that rank highest within a single benchmark are not the models that transfer best. DeepDTA and ESM-DTA collapse to random performance on Davis kinases (per-protein CI 0.520 and 0.501 respectively), despite achieving strong same-dataset metrics on DTC. Anchor transfer reverses this pattern: V2-650M achieves per-protein CI of 0.642 and AUROC of 0.669 on Davis. Crucially, anchors are retrieved at test time from the training set by Tanimoto chemical similarity after excluding all canonical chemical duplicates (83.8% of Davis drugs were present in DTC under different SMILES strings), ensuring genuine zero molecular overlap. Performance stratification by retrieval similarity confirms that the model generalizes even to chemically dissimilar compounds (Tanimoto < 0.6: CI 0.656).

A cosine similarity ablation confirms that the model learns more than protein similarity: 74% of V2-650M's predictive variance is orthogonal to embedding cosine distance, and the full model outperforms cosine similarity on 82% of Davis proteins. The three-way interaction architecture captures drug-specific binding patterns that raw protein similarity cannot encode.

ESM-DTA serves as a critical negative control: it is the strongest same-dataset model (AUROC 0.910 on DTC) yet collapses to random on Davis (per-protein AUROC 0.503). This demonstrates that stronger protein embeddings alone do not solve the cross-dataset transfer problem—the anchor mechanism provides the missing ingredient by conditioning predictions on experimentally grounded binding evidence.

The method is designed for the common drug discovery setting where at least one strong binder of the query drug is already known—lead optimization, scaffold hopping, selectivity profiling, and repurposing all satisfy this condition. Several directions remain open: learned anchor retrieval to improve on the current Tanimoto nearest-neighbor strategy, multi-anchor ensembles for calibrated uncertainty, deeper cross-attention architectures for richer three-way interactions, and evaluation on structurally diverse protein families beyond kinases.

References

[1] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edouard Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao.

Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019. URL https://doi.org/10.1038/s41573-019-0024-5.

[2] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241–1250, 2018. URL https://doi.org/10.1016/j.drudis.2018.01.039.

[3] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018. URL https://doi.org/10.1093/bioinformatics/bty593.

[4] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021. URL https://doi.org/10.1093/bioinformatics/btaa921.

[5] Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021. URL https://doi.org/10.1093/bioinformatics/btaa880.

[6] Rohit Singh, Samuel Sledzieski, Bryan Bryson, Lenore Cowen, and Bonnie Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24):e2220778120, 2023. URL https://doi.org/10.1073/pnas.2220778120.

[7] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014. URL https://doi.org/10.1021/ci400709d.

[8] Mindy I Davis, Jeremy P Hunt, Stephan Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046–1051, 2011. URL https://doi.org/10.1038/nbt.1990.

[9] Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics*, 16(2):325–337, 2015. URL https://doi.org/10.1093/bib/bbu010.

[10] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451, 2018. URL https://doi.org/10.1039/c8sc00148k.

[11] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. URL https://doi.org/10.1109/tkde.2009.191.

[12] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. URL https://doi.org/10.1109/jproc.2020.3004555.

[13] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. URL https://doi.org/10.1126/science.ade2574.

[14] David Weininger. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. URL https://doi.org/10.1021/ci00057a005.

[15] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. Widedta: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*, 2019. URL https://arxiv.org/abs/1902.04166.

[16] Qichang Zhao, Guihua Duan, Mengyun Yang, Zhongjian Cheng, Yaohang Li, and Jianxin Wang. Attentiondta: drug–target binding affinity prediction by sequence-based deep learning with attention mechanism. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):852–863, 2023. URL https://doi.org/10.1109/tcbb.2022.3170365.

[17] Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical Science*, 13(3):816–833, 2022. URL https://doi.org/10.1039/d1sc05180f.

[18] Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug-target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023. doi: 10.1038/s42256-022-00605-1.

[19] Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, 9(1):24, 2017. URL https://doi.org/10.1186/s13321-017-0209-z.

[20] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35 (2):309–318, 2019. URL https://doi.org/10.1093/bioinformatics/bty535.

[21] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. URL https://doi.org/10.1109/tpami.2021.3095381.

[22] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. URL https://doi.org/10.1021/ci100050t.

[23] Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949, 2004. URL https://doi.org/10.1038/nrd1549.

[24] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.01776.

[25] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873): 583–589, 2021. URL https://doi.org/10.1038/s41586-021-03819-2.

[26] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. URL https://doi.org/10.1126/science.abj8754.

[27] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. URL https://doi.org/10.1073/pnas.2016239118.

[28] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html.

[29] Fei Li, Zhaojun Zhang, Jihong Guan, and Shuigeng Zhou. Effective drug–target interaction prediction with mutual interaction neural network. *Bioinformatics*, 38(14):3582–3589, 2022. URL https://doi.org/10.1093/bioinformatics/btac377.

[30] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS Central Science*, 3(4):283–293, 2017. URL https://doi.org/10.1021/acscentsci.6b00367.

[31] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. URL https://doi.org/10.1023/a:1007379606734.

[32] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

[33] Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053, 2016. URL https://doi.org/10.1093/nar/gkv1072.

[34] Mithat Gönen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005. URL https://doi.org/10.1093/biomet/92.4.965.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

[36] Mert Erden, Kapil Devkota, Lia Varghese, Lenore Cowen, and Rohit Singh. Learning a concise language for small-molecule binding. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.632039.

[37] Kapil Devkota, Daichi Shonai, Joey Mao, Scott H. Soderling, and Rohit Singh. Miniaturizing, modifying, and augmenting nature's proteins with raygun. *bioRxiv*, 2024. doi: 10.1101/2024.08.13.607858.

[38] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42:243–246, 2024.

[39] Michael Heinzinger, Konstantin Weissenow, Sofie Kutuzova, and Burkhard Rost. Prostt5: Bilingual language model for protein sequence and structure. *bioRxiv*, 2023. doi: 10.1101/2023.07.23.550085.

[40] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1):41–58, 2019. URL https://doi.org/10.1038/nrd.2018.168.