DELBERT: FINGERPRINT LANGUAGE MODELING FOR GENERALIZABLE HIT DISCOVERY IN DNA-ENCODED LIBRARIES

Arman Seyed-Ahmadi[1], **Bing Hu**[2], **Armin Geraili**[3],
Anita Layton[2], **Helen Chen**[2], **Shana O. Kelley**[3,4], **Bo Wang**[1,3,5]
[1]University Health Network [2]University of Waterloo [3]University of Toronto
[4]Northwestern University [5]Vector Institute
{arman.seyedahmadi, bo.wang}@uhn.ca
{b25hu, anita.layton, helen.chen}@uwaterloo.ca
armin.geraili@utoronto.ca
shana.kelley@northwestern.edu

ABSTRACT

DNA-Encoded Libraries (DEL) enable high-throughput exploration of vast chemical spaces for drug discovery, yet machine learning progress in this domain is limited by the scarcity of publicly available data. Initiatives such as AIRCHECK have begun releasing public DEL data as molecular fingerprints (FPs) only, preserving the confidentiality of proprietary chemical structures while enabling model development. However, current open-source FP-based approaches typically rely on supervised decision tree models that show limited generalization to out-of-distribution (OOD) chemical space. We introduce DELBERT[1], a transformer encoder that treats molecular FPs as a discrete token language, enabling self-supervised pretraining via masked language modeling without requiring access to underlying molecular structures. Under comprehensive library-based OOD evaluation across four protein targets (WDR91, LRRK2, SETDB1, DCAF7), DELBERT significantly outperforms baseline ensemble models on three of four targets, with 1.6–2.7$\times$ improvements in key early-enrichment metrics. Our results demonstrate that self-supervised learning over FPs alone can substantially enhance generalization for hit identification, unlocking confidentiality-preserving collaboration for accelerated drug discovery in data-constrained settings.

1 INTRODUCTION

Hit identification is a critical early step in small-molecule drug discovery, where the goal is to find compounds that bind to a protein target of interest (Iqbal et al., 2025). DNA-Encoded Libraries (DELs) have transformed this process by enabling the simultaneous screening of millions to billions of molecules against a target in a single experiment (Peterson & Liu, 2023). In a DEL screen, combinatorially synthesized molecules are each tagged with a unique DNA barcode; after affinity selection, high-throughput sequencing identifies enriched barcodes, yielding large-scale protein–ligand interaction datasets (Torng et al., 2023). However, translating DEL screen results into validated drug candidates remains challenging: resynthesizing individual compounds off-DNA for confirmatory assays is costly, slow, and often yields low hit rates (Wellnitz et al., 2025; Iqbal et al., 2025). Machine learning (ML) models trained on DEL data offer a compelling alternative, enabling virtual screening (VS) of large external compound libraries to prioritize molecules likely to bind the target, and substantially reducing the time and cost of hit validation (Iqbal et al., 2025).

Several groups have demonstrated the potential of ML for DEL-based hit identification, from the pioneering work of McCloskey et al. (2020) to more recent efforts spanning diverse protein targets and model architectures (Ahmad et al., 2023; Torng et al., 2023; Chen et al., 2024; Ackloo et al., 2025; Wellnitz et al., 2025). Public benchmarks such as BELKA (Quigley et al., 2024) have further

[1]Project repository: https://github.com/bowang-lab/DELBERT

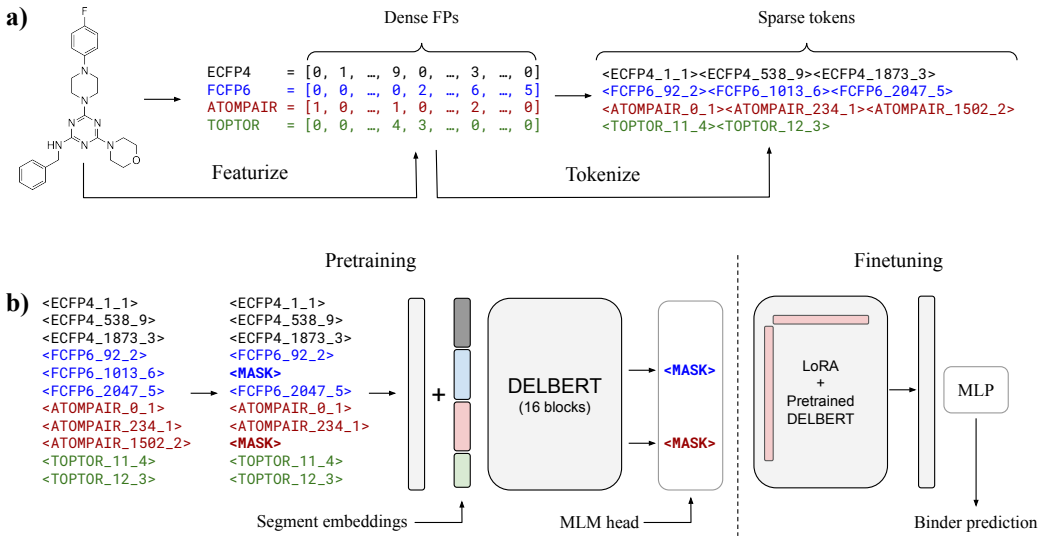


Figure 1: Overview of DELBERT. **(a)** Molecular fingerprints (FPs; ECFP4, FCFP6, ATOMPAIR, TOPTOR) are given as dense count vectors, then tokenized into sparse sequences encoding FP type, bit index, and count. **(b)** During pretraining, tokens are randomly masked and the model learns to reconstruct them via a masked language modeling (MLM) head. Segment embeddings help the model better distinguish FP types. For finetuning, Low-Rank Adaptation (LoRA) adapters are applied to the pretrained encoder and a classification head predicts binding.

catalyzed progress, while also exposing the failure of current models to generalize beyond memorized chemical patterns (Quigley, 2024). Despite these advances, DEL-ML remains a nascent field, largely constrained by the scarcity of publicly available data (Chen et al., 2024); most efforts remain siloed within proprietary pipelines, limiting reproducibility and broad adoption (Barsainyan et al., 2025). The AIRCHECK (Artificial Intelligence-Ready Chemical Knowledge) initiative (Reza et al., 2026; AIRCHECK, 2025) addresses this gap by providing public access to DEL screening data across multiple therapeutically relevant targets. Wellnitz et al. (2025) trained the first successful DEL-ML model on AIRCHECK data, identifying experimentally validated binders via virtual screening. Crucially, however, since DEL providers do not release proprietary library definitions and chemical structures, AIRCHECK can only provide precomputed molecular fingerprints (FPs). This highlights the need for methods that operate effectively on FPs alone until sufficiently large, fully open DEL datasets become available.

This FP-only setting introduces two key challenges. First, it precludes the use of existing molecular encoders that depend on structural information—including graph neural networks applied to DEL hit finding (McCloskey et al., 2020; Torng et al., 2023), self-supervised graph-based approaches (Wang et al., 2022; Zhou et al., 2023), SMILES-based transformers (Chithrananda et al., 2020; Ross et al., 2022), and image- or 3D structure-based representations (Stärk et al., 2022; Zeng et al., 2022). The AIRCHECK DEL screening datasets used in this work are released as precomputed FPs only, making structure-based methods inapplicable. While transformer-based approaches have been explored using Morgan algorithm substructure identifiers as tokens (Méndez-Lucio et al., 2024; Wen et al., 2022), they require access to molecular structures at featurization time. Such methods therefore cannot be applied to the precomputed FP vectors used in this work. Second, existing open-source approaches for DEL hit finding rely on supervised tree-based models (e.g., XGBoost, LightGBM) (Wellnitz et al., 2025). These models can only be trained on labeled data and show limited generalization to novel chemical space (Quigley & Blevins, 2025), as we also confirm in our experiments (Section 3.3).

We introduce DELBERT, a self-supervised transformer encoder designed for representation learning and hit identification using molecular FPs alone. Unlike existing approaches that feed raw FP vectors directly to tree-based classifiers, DELBERT treats each active FP bit as a discrete token, forming a molecular "language" that can be modeled with a transformer. This formulation enables self-supervised pretraining via masked language modeling (MLM) on unlabeled DEL data, followed by parameter-efficient finetuning with LoRA adapters for binding prediction (Figure 1). By first learning the statistical structure of molecular FPs without any binding labels, DELBERT acquires transferable representations that improve downstream generalization.

We evaluate DELBERT on four protein targets from AIRCHECK (WDR91, LRRK2, SETDB1, DCAF7) using a library-based cross-validation protocol where entire chemical libraries are held out at test time. DELBERT significantly outperforms state-of-the-art XGBoost and LightGBM ensembles on three of four targets, with 1.6–2.7$\times$ improvements in early-enrichment metrics. Consistent with recent observations on data leakage in DEL benchmarks (Quigley & Blevins, 2025), we find that tree-based baselines perform well when test molecules are structurally similar to the training set, but degrade substantially on truly out-of-distribution molecules—a setting where DELBERT maintains or even shows improved performance. These results demonstrate that self-supervised learning over FPs alone can yield substantial gains in generalization for DEL hit identification, enabling large-scale, confidentiality-preserving collaboration for drug discovery in data-constrained settings.

2 METHODS

2.1 FINGERPRINT TOKENIZATION

In the HitGen DEL screening datasets from AIRCHECK (Reza et al., 2026) used in this work, molecules are represented by precomputed molecular FPs without access to the underlying chemical structures. DELBERT leverages four complementary FP types: extended-connectivity fingerprints (ECFP4; Rogers & Hahn, 2010), functional-class fingerprints (FCFP6), atom-pair (ATOMPAIR) fingerprints (Carhart et al., 1985), and topological torsion (TOPTOR) fingerprints (Nilakantan et al., 1987). Each FP encodes a different aspect of molecular structure: ECFP4 captures local circular substructures, FCFP6 encodes functional/pharmacophoric features, while atom-pair and topological torsion FPs capture longer-range connectivity patterns.

We propose treating FP bits as discrete tokens to construct a molecular "language" amenable to unsupervised learning. Each active bit in a molecule's FPs is encoded as a token that specifies the FP type, bit position, and occurrence count (e.g., `ECFP4_42_3`). Tokens from all four FP types are concatenated into a single variable-length sequence, forming a holistic molecular representation. A vocabulary is constructed by scanning the pretraining corpus and retaining tokens that appear at least $f_{\min}$ times, with special tokens `[MASK]` and `[PAD]` added for model use.

2.2 MODEL ARCHITECTURE

DELBERT is built on ModernBERT (Warner et al., 2024), an efficiency-optimized BERT variant featuring Flash Attention (Dao et al., 2022) with unpadding, rotary position embeddings (RoPE; Su et al., 2024), and alternating local-global attention layers. Local layers use a sliding window for efficient processing, while periodic global layers allow information flow across the full sequence. The model uses a hidden dimension of 640, 16 transformer layers, 16 attention heads, and a feed-forward intermediate size of 1,152, totaling 70M parameters, with a maximum sequence length of 1,024 tokens. Pretraining on $\sim 500K$ molecules takes approximately 10 hours and finetuning approximately 20 minutes on $4\times$H100 80GB GPUs. For classification, we apply mean pooling over non-padding token representations followed by dropout and a linear projection to predict binary binding labels.

2.3 SELF-SUPERVISED PRETRAINING

DELBERT is pretrained using masked language modeling (MLM) on DEL data without using binding labels. For each molecule, 15% of tokens are randomly selected for masking: 80% are replaced with `[MASK]`, 10% with a random token, and 10% are left unchanged. The model is trained to predict the original token at each masked position via cross-entropy loss. Pretraining uses all molecules

from the target's DEL screen (both enriched and non-enriched), discarding binding labels entirely. Pretraining is performed separately for each protein target rather than pooling across targets, since the four AIRCHECK targets share substantially overlapping chemical libraries (Figure A.1). We use AdamW with learning rate 5×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.98)$, weight decay 0.01, and cosine schedule with 10% linear warmup for 100 epochs.

2.4 Supervised Finetuning

For binding prediction, we finetune the pretrained model using Low-Rank Adaptation (LoRA; Hu et al., 2022) with rank r=16, scaling factor α=32, and dropout 0.1, applied to the query-key-value attention projections. This approach keeps the pretrained encoder largely frozen while adapting through low-rank updates, providing implicit regularization against overfitting. We use weighted cross-entropy loss with a positive class weight of 15.0 to address the severe class imbalance in DEL data (5–11% positive rate). Training runs for up to 10 epochs with early stopping on PR-AUC computed on a random validation split within each fold's training data (patience of 5). Notably, we use the same hyperparameters across all protein targets and folds without any target-specific tuning to avoid overfitting to particular evaluation settings.

2.5 Experimental Setup

Datasets. We evaluate on four protein targets from HitGen's DEL screening data in AIRCHECK (Reza et al., 2026): WDR91 (376K molecules), LRRK2 (321K), SETDB1 (420K), and DCAF7 (512K). Each dataset comprises multiple chemical libraries screened against a protein target, with binary labels indicating enrichment (see Figure A.1 for detailed dataset composition).

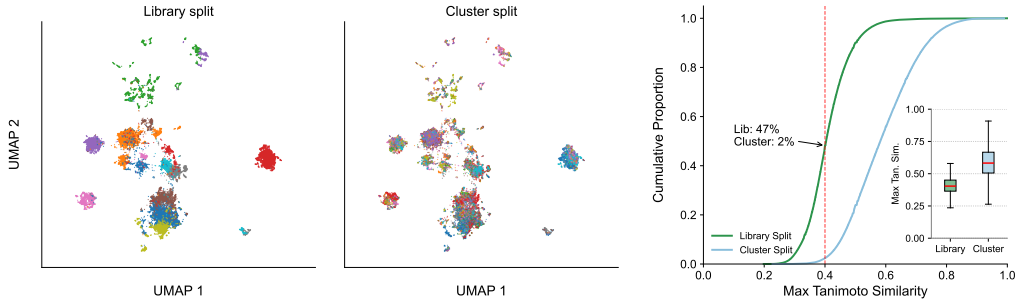
Evaluation. Prior work on DEL ML has typically used cluster-based or random splitting for evaluation (Wellnitz et al., 2025). However, clustering could produce groups that intermix across the chemical space of different libraries, leaving structurally similar molecules in both train and test sets and inflating performance estimates (see Section 3.1). Following McCloskey et al. (2020), we adopt library-based K-fold cross-validation, where entire chemical libraries are held out per fold to provide a more rigorous OOD evaluation. That is, molecules in the test fold come from libraries completely absent during training. Libraries with fewer than 1,000 molecules are grouped and kept in the training set. In this setting, K is determined by each target's library composition, yielding $K \in \{11, 15, 18, 20\}$ depending on the dataset. We report ROC-AUC, PR-AUC, BEDROC (Truchon & Bayly, 2007) (α=85), Precision@100, and Enrichment Factor@100 (EF@100). To assess statistical significance, we apply the Wilcoxon signed-rank test ($p < 0.05$) over fold-level metric values between each pair of models, testing whether one model consistently outperforms the other across folds. To further probe OOD generalization, we stratify test molecules by maximum Tanimoto similarity to the nearest training molecule ($Tc_{\max}$), using a threshold of 0.4 (Steshin, 2023) to separate structurally similar from truly novel compounds, and report metrics on each subset separately.

Baselines. As the only published open-source method for DEL hit finding from precomputed FPs, we use the ensemble approach of Wellnitz et al. (2025) as our baseline. We employ both XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), each using 300 estimators with default hyperparameters. For each outer CV fold, the training set is partitioned into 5 stratified groups using LeaderPicker clustering on binarized FCFP6 FPs (Tanimoto distance threshold of 0.8), and a separate model is trained on each group for each of the four FP types, yielding 20 models per fold. Final predictions are obtained by averaging the predicted probabilities across all models. Baselines use identical outer fold assignments and the same four FP types (binarized) as DELBERT for fair comparison.

3 Results

3.1 Library-based Splits

A key motivation for our evaluation strategy is that cluster-based splits may not adequately separate structurally similar molecules across train and test sets. Figure 2a illustrates this on WDR91: a



(a) (b)

Figure 2: Effect of splitting strategy on OOD evaluation. **(a)** UMAP of WDR91 FP space (FCFP6) colored by fold assignment: library splits produce separated folds, while cluster splits intermix across chemical space. **(b)** CDF of maximum Tanimoto similarity to the nearest training molecule for each test molecule. The dashed red line marks the 0.4 OOD threshold (Steshin, 2023): 47% of library-split test molecules can be considered novel vs. only 2% for cluster splits.

UMAP projection of the molecular FP space colored by library membership shows that library-based folds produce well-separated clusters, whereas cluster-based folds (LeaderPicker on FCFP6, threshold 0.85) exhibit extensive intermixing—molecules from many different folds co-occur in the same regions of chemical space. Notably, our library-based splits produce a separation in UMAP space analogous to the UMAP-based clustering split recently proposed by Guo et al. (2025) for rigorous virtual screening evaluation, arising naturally from the experimental library structure rather than requiring a post-hoc clustering step.

To quantify this, we compute the maximum Tanimoto similarity between each test molecule and its nearest neighbor in the training set across all folds (Figure 2b). Using the 0.4 Tanimoto similarity threshold proposed by Steshin (2023) as a boundary between in-distribution and OOD molecules, 47% of test molecules in the library split fall below this threshold (i.e., are truly OOD), compared to only 2% in the cluster split. This demonstrates that cluster-based splitting leaves the vast majority of test molecules with highly similar training neighbors, substantially overestimating model generalization (Figure A.2 further illustrates this by showing how test molecules form a distinct island under library splits, but overlap significantly with the training molecules under cluster splits). All results in this work are therefore based on the library-based splitting.

3.2 Model Comparison

Table 1 and Figure 3 summarize the library-based cross-validation results across all four protein targets. DELBERT significantly outperforms both XGBoost and LightGBM ensembles on WDR91, LRRK2, and SETDB1 across all five evaluation metrics, with all differences statistically significant by the Wilcoxon signed-rank test ($p < 0.01$ or $p < 0.001$). On DCAF7, all three models perform comparably, with no statistically significant differences.

The gains are most pronounced in early-enrichment metrics. These metrics are the most practically relevant for virtual screening, where only a small number of top-ranked compounds can be experimentally validated. Compared to the best baseline, DELBERT achieves 1.6×, 2.1×, and 2.4× higher BEDROC on WDR91, LRRK2, and SETDB1 respectively, with corresponding improvements in EF@100 (1.9×, 2.6×, 2.7×) and Precision@100 (from 0.34 to 0.56, 0.20 to 0.46, and 0.28 to 0.73). The largest improvements are on SETDB1, where ROC-AUC increases from 0.71 to 0.92 and BEDROC from 0.30 to 0.71. DCAF7 is the exception, with all three models performing comparably at ∼0.86 ROC-AUC; the target-specific factors underlying this result warrant further investigation. Supplementary ROC and PR curves (Figure A.4) and fold-level paired comparisons (Figure A.3) are provided in the Appendix.

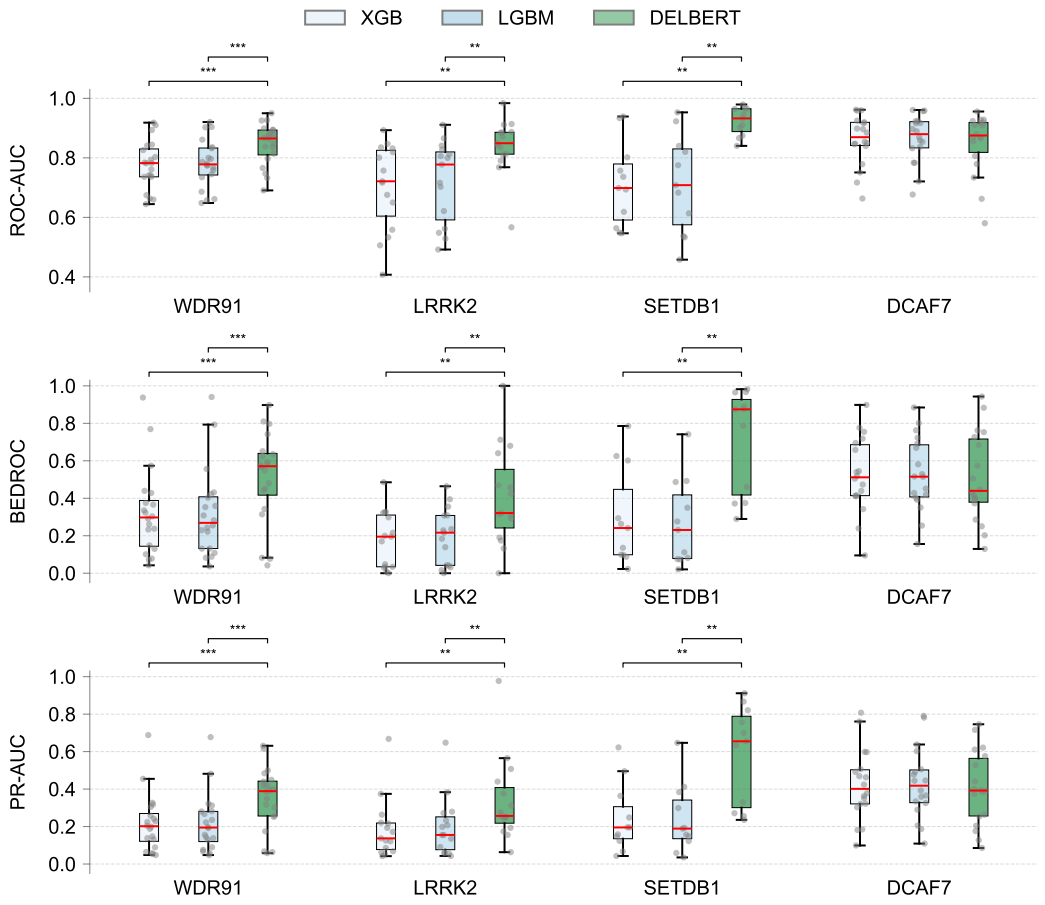


Figure 3: Distribution of fold-level ROC-AUC, BEDROC, and PR-AUC across four protein targets for XGBoost, LightGBM, and DELBERT. Each point is a single CV fold; boxes span the interquartile range (IQR) with whiskers extending to 1.5× IQR; red lines show the median. Significance brackets indicate Wilcoxon signed-rank test results (** $p < 0.01$, *** $p < 0.001$). DELBERT achieves significantly higher performance on WDR91, LRRK2, and SETDB1 across all metrics.

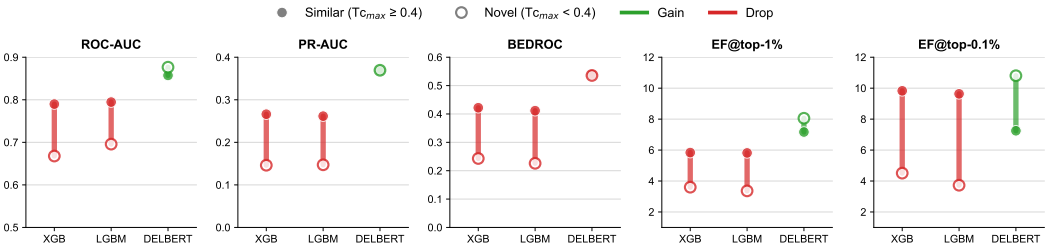


Figure 4: Performance on similar ($Tc_{\max} \geq 0.4$) vs. novel ($Tc_{\max} < 0.4$) test molecules on WDR91, where $Tc_{\max}$ is the maximum Tanimoto similarity to the nearest training molecule. Filled circles denote similar molecules, whereas open circles denote novel/OOD molecules. Green bars indicate a gain and red bars a drop when moving from similar to novel. XGBoost and LightGBM show consistent drops across all metrics, while DELBERT maintains or shows better performance on OOD molecules, demonstrating stronger generalization to novel chemical space.

Table 1: Library-based cross-validation results comparing XGBoost, LightGBM, and DELBERT across four protein targets. Values are mean ± std (median) over K folds (number of folds varies by target due to differing library compositions). **Bold** indicates best performance when statistically significant (Wilcoxon signed-rank test, $p < 0.05$); underline indicates second best when significantly better than third. DELBERT significantly outperforms both baselines on WDR91, LRRK2, and SETDB1 across all metrics, with similar performance on DCAF7.

Metric	Model	WDR91 (K=20)	LRRK2 (K=15)	SETDB1 (K=11)	DCAF7 (K=18)
ROC-AUC	XGB	0.777 ± 0.079 (0.782)	0.703 ± 0.141 (0.721)	0.713 ± 0.134 (0.699)	0.859 ± 0.083 (0.869)
	LGBM	0.781 ± 0.078 (0.778)	0.721 ± 0.132 (0.778)	0.713 ± 0.157 (0.708)	0.864 ± 0.079 (0.880)
	DELBERT	**0.846 ± 0.068 (0.865)**	**0.841 ± 0.090 (0.849)**	**0.923 ± 0.045 (0.933)**	0.850 ± 0.099 (0.875)
BEDROC	XGB	0.322 ± 0.225 (0.298)	0.180 ± 0.145 (0.195)	0.296 ± 0.246 (0.241)	0.524 ± 0.195 (0.512)
	LGBM	0.319 ± 0.227 (0.269)	0.198 ± 0.145 (0.216)	0.267 ± 0.219 (0.231)	0.529 ± 0.190 (0.515)
	DELBERT	**0.513 ± 0.236 (0.571)**	**0.408 ± 0.253 (0.321)**	**0.712 ± 0.264 (0.875)**	0.509 ± 0.230 (0.440)
PR-AUC	XGB	0.218 ± 0.150 (0.202)	0.187 ± 0.156 (0.137)	0.242 ± 0.173 (0.195)	0.423 ± 0.183 (0.401)
	LGBM	0.219 ± 0.150 (0.195)	0.198 ± 0.153 (0.155)	0.240 ± 0.173 (0.189)	0.430 ± 0.183 (0.418)
	DELBERT	**0.349 ± 0.162 (0.389)**	**0.336 ± 0.215 (0.257)**	**0.585 ± 0.250 (0.655)**	0.405 ± 0.192 (0.392)
Prec@100	XGB	0.345 ± 0.258 (0.305)	0.190 ± 0.190 (0.150)	0.283 ± 0.237 (0.240)	0.561 ± 0.209 (0.545)
	LGBM	0.336 ± 0.261 (0.275)	0.202 ± 0.167 (0.200)	0.254 ± 0.236 (0.150)	0.551 ± 0.210 (0.510)
	DELBERT	**0.563 ± 0.265 (0.590)**	**0.457 ± 0.291 (0.390)**	**0.733 ± 0.264 (0.890)**	0.511 ± 0.244 (0.445)
EF@100	XGB	5.621 ± 3.800 (4.427)	2.598 ± 2.154 (1.762)	3.343 ± 3.008 (2.903)	8.669 ± 6.618 (6.053)
	LGBM	5.512 ± 3.849 (4.812)	2.887 ± 2.153 (1.683)	2.988 ± 2.847 (2.349)	8.614 ± 6.732 (5.991)
	DELBERT	**10.456 ± 8.075 (8.194)**	**7.428 ± 5.358 (6.555)**	**8.985 ± 4.071 (9.594)**	7.046 ± 4.596 (5.740)

3.3 OOD Robustness

To understand why DELBERT's advantage is greatest under library-based splits, we analyze performance separately on structurally similar ($Tc_{\max} \geq 0.4$) and novel ($Tc_{\max} < 0.4$) test molecules on WDR91, as shown in Figure 4. This threshold coincides with the median $Tc_{\max}$ across test molecules, yielding two equally sized subsets and avoiding artifacts from imbalanced group sizes. XGBoost and LightGBM show consistent performance drops when evaluated on novel molecules across all metrics, indicating that their predictions rely heavily on structural similarity to the training set. In contrast, DELBERT maintains or even shows improved performance on novel molecules, suggesting that its pretrained representations capture generalizable molecular patterns beyond surface-level structural similarity. This is consistent with Quigley & Blevins (2025), who showed that pretrained models only surpass XGBoost as train–test structural overlap decreases, with supervised models degrading sharply on molecules sharing no building blocks with the training set. DELBERT's robustness to novel chemical space directly explains its larger gains under library-based splits, where nearly half of test molecules can be considered OOD.

3.4 Ablations

We conduct an ablation study on WDR91 (20-fold library CV) to quantify the contribution of each design choice (Table 2). Randomly initializing the model and training it directly on the labeled data results in the largest performance drop (BEDROC: 0.513 → 0.249; ROC-AUC: 0.846 → 0.678), confirming that self-supervised pretraining is the most critical component. Using only ECFP4 instead of four FP types causes the second-largest drop (BEDROC: 0.361), demonstrating the value of multi-FP input for capturing complementary molecular features. Reducing model capacity (17M parameters) and binarizing FP counts both degrade performance moderately (BEDROC: 0.366 and 0.429 respectively), indicating that model expressiveness and count-based tokenization each contribute meaningfully. The latter is notable, since using count-based FPs provided no improvement for baseline models, whereas DELBERT appears to be able to utilize the richer representation for better predictions. Adding more FP types (7 instead of 4, including ECFP6, FCFP4, and MACCS (Durant et al., 2002)) provides marginal gains on some metrics (BEDROC: 0.455) but not others,

Table 2: Ablation study on WDR91 (20-fold library-based CV). Each row removes or modifies one component of DELBERT: pretraining, multi-FP input, model capacity, or count-based tokenization. Values are mean ± std (median). **Bold** indicates best; underline second best. All components contribute to the full model's performance, with pretraining and multi-FP input having the largest impact.

Ablation	ROC-AUC	BEDROC	PR-AUC	Prec@100
No pretraining	0.678 ± 0.099 (0.709)	0.249 ± 0.208 (0.224)	0.158 ± 0.116 (0.134)	0.265 ± 0.232 (0.235)
ECFP4 only	0.803 ± 0.074 (0.820)	0.361 ± 0.194 (0.364)	0.251 ± 0.128 (0.259)	0.371 ± 0.228 (0.355)
Smaller model	0.806 ± 0.067 (0.817)	0.366 ± 0.220 (0.368)	0.252 ± 0.146 (0.262)	0.385 ± 0.242 (0.415)
Binary FPs	0.834 ± 0.059 (0.848)	0.429 ± 0.228 (0.432)	0.295 ± 0.151 (0.291)	0.456 ± 0.259 (0.455)
More FPs	0.832 ± 0.054 (0.836)	0.455 ± 0.171 (0.468)	0.302 ± 0.130 (0.336)	0.490 ± 0.222 (0.475)
DELBERT	**0.846 ± 0.068 (0.865)**	**0.513 ± 0.236 (0.571)**	**0.349 ± 0.162 (0.389)**	**0.563 ± 0.265 (0.590)**

suggesting diminishing returns beyond four complementary types. Overall, pretraining and multi-FP input seem to be the two most impactful design choices in DELBERT.

4 Limitations and Future Work

DEL molecules are assembled combinatorially from a limited set of building blocks (typically several hundred to a thousand), which constrains the chemical diversity that DELBERT can learn from during pretraining (Gu et al., 2024). Additionally, DEL data is inherently noisy: library synthesis is optimized for scale rather than precision, producing partial or truncated products, and the final readout relies on PCR amplification of DNA barcodes, which does not uniformly sample the surviving library members (Chen et al., 2024). While DELBERT's self-supervised pretraining helps extract signal from this noisy data, these factors may still limit the quality of learned representations. Furthermore, our evaluation is currently limited to retrospective DEL-to-DEL prediction; we have not yet validated DELBERT's ability to identify hits in external non-DEL compound collections, which is the ultimate goal for virtual screening applications. Since the AIRCHECK datasets used in this work provide FPs only, our comparisons are necessarily limited to FP-compatible methods; evaluating DELBERT against structure-based approaches (GNNs, SMILES transformers) on datasets where molecular structures are available (Chen et al., 2024) remains an important direction for follow-up work.

We also note that while molecular fingerprints can in principle be reverse-engineered (Meyer et al., 2025), AIRCHECK's data sharing agreements with industry partners explicitly prohibit determining chemical structures from the provided FPs (AIRCHECK, 2025); nonetheless, the privacy implications of releasing multiple FP types warrant further investigation. Finally, although DEL chemical space is constrained by combinatorial synthesis, the FP co-occurrence patterns captured by MLM still encode genuine molecular substructure relationships, providing a useful inductive bias for downstream binding prediction.

Moreover, a natural next step is to pretrain DELBERT on large external molecular databases such as ZINC (Sterling & Irwin, 2015), PubChem (Kim et al., 2023), and ChEMBL (Gaulton et al., 2012)—datasets routinely used by molecular foundation models (Ross et al., 2022; Chithrananda et al., 2020; Zhou et al., 2023)—to infuse the model with broader chemical knowledge beyond DEL space. This could be particularly impactful when screening external compound catalogues where DEL-only pretraining may be insufficient. Looking further ahead, we plan to use DELBERT to screen large-scale commercial catalogues, nominate candidate compounds, and pursue experimental validation through biophysical assays such as surface plasmon resonance (SPR).

5 Conclusion

We presented DELBERT, a self-supervised transformer encoder that reframes molecular fingerprints as a discrete token language, enabling masked language modeling on DEL data without the need to access underlying chemical structures. Our central finding is that self-supervised pretraining on FPs alone—usually considered too lossy for representation learning—yields representations

that substantially outperform state-of-the-art supervised ensembles, with 1.6–2.7× improvements in early-enrichment metrics across three of four protein targets. Crucially, the advantage is not uniform: it emerges most clearly under rigorous OOD evaluation, where baselines degrade on structurally novel molecules while DELBERT maintains or shows improved performance. This suggests that supervised decision tree-based models, despite strong in-distribution performance, largely exploit surface-level similarity to training compounds, whereas self-supervised pretraining learns more transferable molecular representations. Ablation analysis confirms that pretraining is the most important component, and more broadly, our results underscore the need for realistic OOD evaluations, without which the generalization gaps may remain hidden. From a practical standpoint, DELBERT directly addresses a key bottleneck in DEL-based drug discovery: it enables effective virtual screening models to be built on privacy-preserving FP-only data, lowering the barrier for cross-institutional collaboration in data-constrained therapeutic programs.

ACKNOWLEDGMENTS

We thank Luca Chiesa, Shaghayegh Reza, and the AIRCHECK team for valuable discussions, and Benjamin Haibe-Kains for kindly providing insights and detailed feedback on the manuscript. Computational resources were provided by the Vector Institute and the Digital Research Alliance of Canada (`https://alliancecan.ca`).

REFERENCES

Suzanne Ackloo, Fengling Li, Magda Szewczyk, Almagul Seitova, Peter Loppnau, Hong Zeng, Jin Xu, Shabbir Ahmad, Yelena A Arnautova, AJ Bagale, et al. A target class ligandability evaluation of wd40 repeat-containing proteins. *Journal of Medicinal Chemistry*, 68(2):1092–1112, 2025.

Shabbir Ahmad, Jin Xu, Jianwen A Feng, Ashley Hutchinson, Hong Zeng, Pegah Ghiabi, Aiping Dong, Paolo A Centrella, Matthew A Clark, Marie-Aude Guié, et al. Discovery of a first-in-class small-molecule ligand for wdr91 using dna-encoded chemical library selection followed by machine learning. *Journal of Medicinal Chemistry*, 66(23):16051–16061, 2023.

AIRCHECK. Artificial Intelligence-Ready CHEmiCal Knowledge Base (AIRCHECK). `https://aircheck.ai`, 2025.

Aryan Amit Barsainyan, Riya Singh, Abhiraj Pravin Mengade, Rida Irfan, and Bharath Ramsundar. Deepchem-del: An open source framework for reproducible del modeling and benchmarking, 2025. ChemRxiv preprint.

Raymond E Carhart, Dennis H Smith, and RENGACHARI Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.

Benson Chen, Tomasz Danel, Gabriel HS Dreiman, Patrick J McEnaney, Kirill Novikov, Spurti Umesh Akki, Joshua L Turnbull, Virja Atul Pandya, Boris P Belotserkovskii, et al. Kindel: Dna-encoded library dataset for kinase inhibitors. *arXiv preprint arXiv:2410.08938*, 2024.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Chunbin Gu, Mutian He, Hanqun Cao, Guangyong Chen, Chang-yu Hsieh, and Pheng Ann Heng. Unlocking potential binders: Multimodal pretraining del-fusion for denoising dna-encoded libraries. *arXiv preprint arXiv:2409.05916*, 2024.

Qianrong Guo, Saiveth Hernandez-Hernandez, and Pedro J Ballester. UMAP-based clustering split for rigorous evaluation of AI models for virtual screening on cancer cell lines. *Journal of Cheminformatics*, 17(1):94, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Sumaiya Iqbal, Wei Jiang, Eric Hansen, Tonia Aristotelous, Shuang Liu, Andrew Reidenbach, Cerise Raffier, Alison Leed, Chengkuan Chen, Lawrence Chung, et al. Evaluation of dna encoded library and machine learning model combinations for hit discovery. *npj Drug Discovery*, 2 (1):5, 2025.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

Kevin McCloskey, Eric A Sigel, Steven Kearnes, Ling Xue, Xia Tian, Dennis Moccia, Diana Gikunju, Sana Bazzaz, Betty Chan, Matthew A Clark, et al. Machine learning on dna-encoded libraries: a new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866, 2020.

Oscar Méndez-Lucio, Christos A Nicolaou, and Berton Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1):9431, 2024.

Philippe Meyer, Thomas Duigou, Guillaume Gricourt, and Jean-Loup Faulon. Reverse engineering molecules from fingerprints through deterministic enumeration and generative models. *Journal of Cheminformatics*, 17(1):157, 2025. doi: 10.1186/s13321-025-01074-5.

Ramaswamy Nilakantan, Norman Bauman, J Scott Dixon, and R Venkataraghavan. Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*, 27(2):82–85, 1987.

Alexander A Peterson and David R Liu. Small-molecule discovery through dna-encoded libraries. *Nature Reviews Drug Discovery*, 22(9):699–722, 2023.

Ian Quigley. BELKA results suggest computers can memorize, but not create, drugs, aug 2024. URL `https://leashbio.substack.com/p/belka-results-suggest-computers-can`. Leash Substack blog post.

Ian Quigley and Andrew Blevins. Data contamination is all random forest needs, jul 2025. URL `https://leashbio.substack.com/p/data-contamination-is-all-random`. Leash Substack blog post.

Ian K Quigley, Andrew Blevins, Brayden J Halverson, and Nate Wilkinson. Belka: The big encoded library for chemical assessment. In *NeurIPS 2024 Competition Track*, 2024.

Shaghayegh Reza, Nabin Bagale, James Wellnitz, Luca Chiesa, Rafael M. Couñago, Sofia Melliou, Claudia Gordijo, Seth Veenbaas, Jonathan Chapman, Valeriia Kaneva, Sisira k. Nair, Kiana Jenabidehkordi, Jianxian Sun, Xiaoyun Wang, Cheryl Arrowsmith, Aled Edwards, Levon Halabelian, Hui Peng, Albert A. Antolin, Matthieu Schapira, Alexander Tropsha, and Benjamin Haibe-Kains. Aircheck: An open platform for ai-driven drug discovery. *ChemRxiv*, 2026(0331), 2026. doi: 10.26434/chemrxiv.15001454/v1. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv.15001454/v1.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3D infomax improves GNNs for molecular property prediction. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20479–20502, 2022.

Teague Sterling and John J. Irwin. ZINC 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559.

Simon Steshin. Lo-hi: Practical ML drug discovery benchmark. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Wen Torng, Ilaria Biancofiore, Sebastian Oehler, Jin Xu, Jessica Xu, Ian Watson, Brenno Masina, Luca Prati, Nicholas Favalli, Gabriele Bassi, et al. Deep learning approach for the discovery of tumor-targeting small organic ligands from dna-encoded chemical libraries. *ACS omega*, 8(28): 25090–25100, 2023.

Jean-François Truchon and Christopher I Bayly. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling*, 47(2):488–508, 2007.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663.

James Wellnitz, Shabbir Ahmad, Nabin Bagale, Xuemin Cheng, Jermiah Joseph, Hong Zeng, Albina Bolotokova, Aiping Dong, Shaghayegh Reza, Pegah Ghiabi, et al. Enabling open machine learning of deoxyribonucleic acid-encoded library selections to accelerate the discovery of small molecule protein binders. *Journal of Medicinal Chemistry*, 68(20):21635–21648, 2025.

Naifeng Wen, Guanqun Liu, Jie Zhang, Rubo Zhang, Yating Fu, and Xu Han. A fingerprints based molecular property prediction method using the BERT model. *Journal of Cheminformatics*, 14 (1):71, 2022.

Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The eleventh international conference on learning representations*, 2023.

A APPENDIX

A.1 DATASET COMPOSITION

Figure A.1 shows the composition of each target's DEL data. Library sizes vary widely both within and across targets, and class imbalance is severe (5–11% hit rate). Libraries with fewer than 1,000 molecules are grouped into a permanent training partition.

A.2 TRAIN–TEST OVERLAP UNDER DIFFERENT SPLITS

Figure A.2 visualizes the train and test sets for a single fold under library-based vs. cluster-based splitting on WDR91. Under the library split, the test set forms a clearly separated island in UMAP space, confirming that held-out libraries occupy distinct regions of chemical space. Under the cluster split, test molecules are dispersed throughout the training distribution, resulting in extensive structural overlap between train and test sets.

A.3 FOLD-LEVEL PAIRED COMPARISONS

Figure A.3 provides a fold-level paired comparison between LightGBM (the stronger baseline) and DELBERT. Each line connects the same fold's scores for the two models, making it easy to see whether improvements are consistent or driven by a few outlier folds. DELBERT outperforms LightGBM on the majority of folds for WDR91, LRRK2, and SETDB1, while the two models perform similarly on DCAF7 (i.e. no statistically significant differences are observed).

A.4 ROC AND PRECISION–RECALL CURVES

Figure A.4 shows mean ROC and Precision–Recall curves across all CV folds. DELBERT consistently achieves a higher area under both curves on WDR91, LRRK2, and SETDB1, with the largest separation visible in the high-precision/low-recall region of the PR curves, which is the operating regime most relevant to virtual screening.

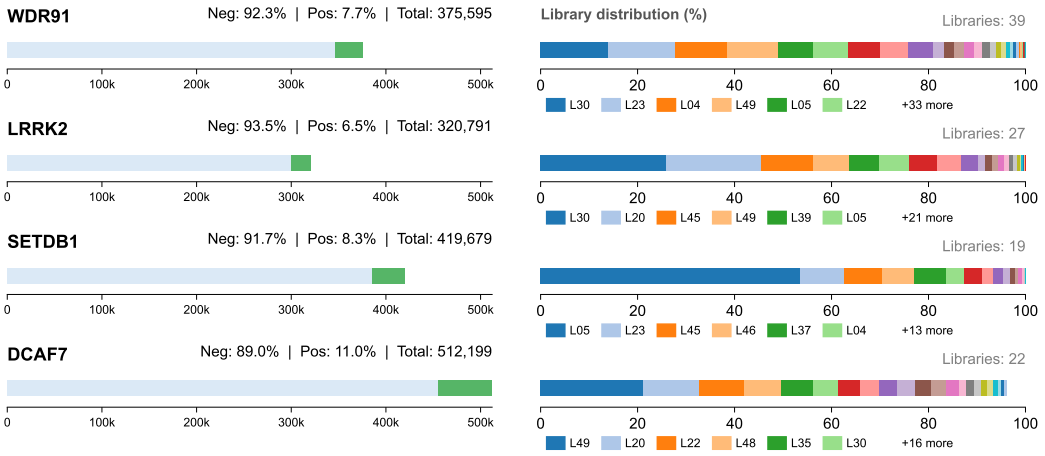


Figure A.1: Dataset composition for each protein target. Bar heights show the number of molecules per chemical library, colored by enrichment label (hit vs. non-hit). Library sizes vary widely both within and across targets.

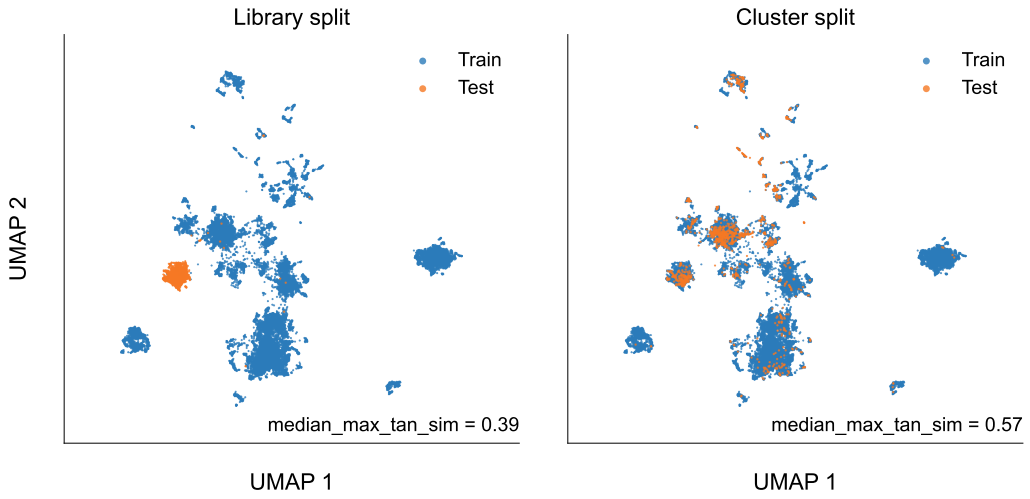


Figure A.2: UMAP visualization of train (blue) and test (orange) molecules for a single fold under library-based vs. cluster-based splitting on WDR91. In the library split, test molecules occupy a distinct region of chemical space (median max Tanimoto similarity to training set = 0.39). In the cluster split, test molecules are dispersed throughout the training distribution (median max Tanimoto similarity = 0.57), indicating substantial structural overlap between train and test sets.

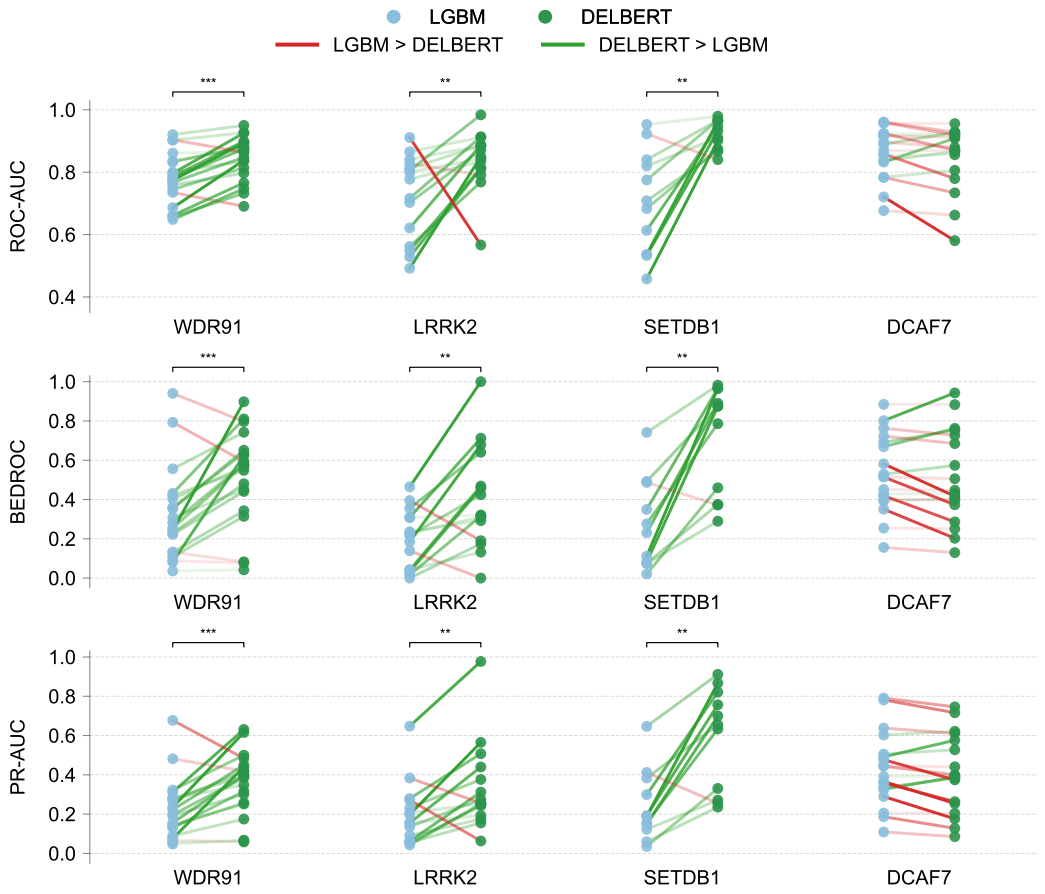


Figure A.3: Fold-level paired comparison of LightGBM and DELBERT across four protein targets for ROC-AUC, BEDROC, and PR-AUC. Each line connects the same fold's scores for the two models; green indicates DELBERT outperforms LightGBM, red the opposite. DELBERT improves over LightGBM on the majority of folds across most targets and metrics.

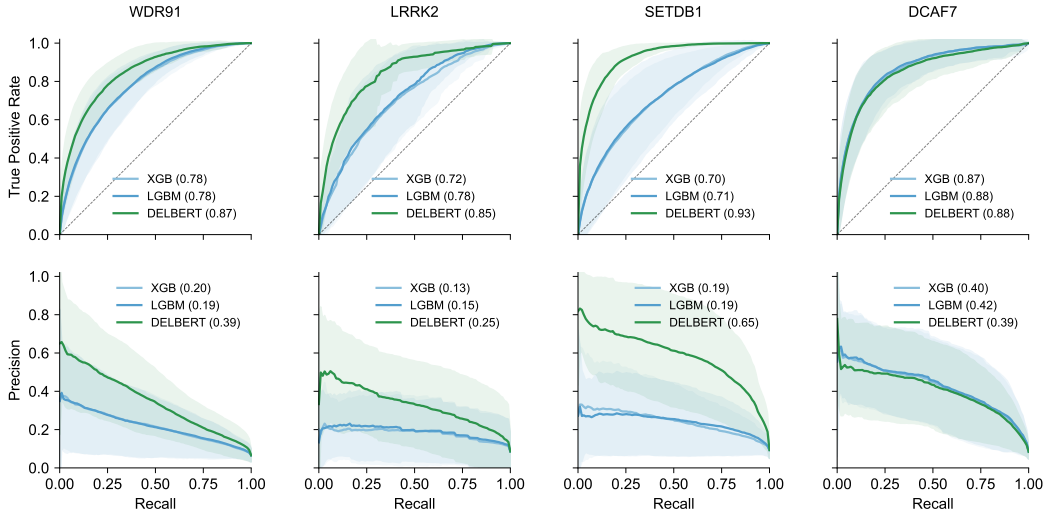


Figure A.4: Mean ROC curves (top) and Precision-Recall curves (bottom) across CV folds for all four protein targets. Shaded regions indicate ± 1 standard deviation across folds. Legend values show mean AUC. DELBERT achieves higher ROC-AUC and PR-AUC than both baselines on WDR91, LRRK2, and SETDB1, with comparable performance on DCAF7.