DELBERT-2: Pretrained Fingerprint Language Models for DEL Protein Binder Prediction

Bing Hu [1] **Sun Sun** [2] **Shaik Salman Basha** [1] **Anita Layton** [3] **Helen Chen** [4]

Abstract

We present DELBERT-2, a pretrained fingerprint language model that converts sparse molecular fingerprints into unified token sequences using a ModernBERT encoder trained with masked language modelling on 2.5M molecules from the AIRCHECK DEL corpus. We evaluate across six targets (WDR91, WDR12, SETDB1, PLCZ1, LRRK2, DCAF7) under three out-of-distribution (OOD) protocols: hierarchical cluster splits probe chemical novelty, library splits test cross-library transfer, and building-block splits assess compositional generalization. DELBERT-2 consistently improves PR-AUC and NDCG@1000 relative to LightGBM ensemble baselines and transformers trained from scratch, with the largest gains in stringent OOD regimes. DELBERT-2 achieves 13.28±3.72 enrichment factor at top-100 vs. 12.36±3.75 for no-pretraining (p=0.040), representing 7.4% improvement and 13× enrichment over random selection. These results demonstrate that fingerprint-centric self-supervised learning effectively improves hit prioritization under distribution shift, enabling practical DEL virtual screening. Code is available at DELBERT-2.

1. Introduction

Hit identification remains a major bottleneck in early-stage drug discovery: the goal is to find small molecules that bind a desired protein target with sufficient quality to justify downstream optimization (Wellnitz et al., 2024). For many years, high-throughput screening using functional assays has been the predominant approach for discovering small-molecule hits (Blay et al., 2020). One of the more cost-effective technologies to emerge is the DNA-encoded library (DEL), in which each compound is tagged with a unique DNA barcode that encodes its synthetic history and composition (Brenner & Lerner, 1992). DELs facilitate the rapid screening of billions of compounds at relatively low cost (Gironda-Martínez et al., 2021).

At the same time, DEL screening has produced increasingly large public datasets, including those released through the Artificial Intelligence Ready Chemical Knowledge (AIRCHECK) database (Reza et al., 2026). These resources enable machine learning models for DEL virtual screening, prioritizing candidates for experimental follow-up without costly resynthesis. However, data scarcity remains a fundamental challenge (Mswahili & Jeong, 2024; Wellnitz et al., 2024), and current DEL-ML methods often fail on compounds that differ substantially from the training distribution—including molecules from unseen libraries, unseen building blocks, or chemically novel regions of space (Quigley et al., 2024; Wellnitz et al., 2024). Improving out-of-distribution (OOD) generalization is therefore critical for making DEL-ML practically useful.

Transformers have emerged as a powerful architecture for learning self-supervised representations, where pretraining followed by task-specific fine-tuning can provide substantial gains on downstream tasks, especially under distribution shift (Hendrycks et al., 2020). In molecular machine learning, much of this progress has focused on SMILES strings or graph representations. However, in practice, many strong DEL screening pipelines continue to rely on molecular fingerprints because they are efficient, chemically informative, and widely used in both public and industrial workflows (Cereto-Massagué et al., 2015; Mswahili & Jeong, 2024). Furthermore, public DEL datasets such as AIRCHECK release only precomputed molecular fingerprints to preserve the confidentiality of proprietary chemical structures, making fingerprint-only approaches essential for broad collaboration.

Recent work explored fingerprint language modelling for DEL hit prediction (Seyed-Ahmadi et al., 2026). Building on this, we introduce DELBERT-2, which extends the framework through *unified pre-training* on the entire AIRCHECK DEL corpus (2.5M molecules vs. 500K per-target) and

[1]Computer Science, University of Waterloo, Canada [2]National Research Council of Canada, Canada [3]Applied Mathematics, University of Waterloo, Canada [4]School of Public Health, University of Waterloo, Canada. Correspondence to: Bing Hu <b25hu@uwaterloo.ca>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

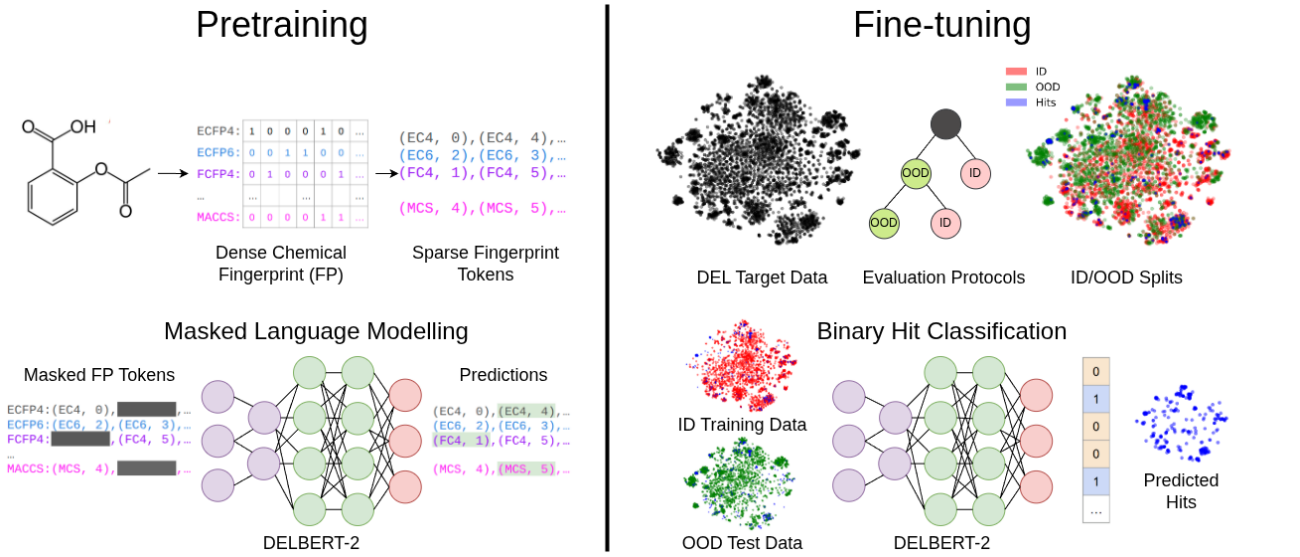


Figure 1. DELBERT-2 pre-training and fine-tuning workflow. Multiple sparse fingerprints are tokenized into a unified sequence: each active bit at position `pos` in fingerprint type `fp` becomes a discrete token `fp_pos_val`, transforming sparse binary vectors into variable-length token sequences. The model is pre-trained with masked language modeling (MLM) on 2.5M DEL molecules, then fine-tuned with LoRA for target-specific binder prediction.

comprehensive OOD evaluation across three complementary protocols (Figure 1). Across six AIRCHECK targets, DELBERT-2 consistently improves PR-AUC and NDCG@1000 relative to strong LightGBM baselines and transformers trained from scratch, with the clearest gains in more stringent OOD regimes.

Our contributions are as follows:

- We demonstrate that unified pretraining on 2.5M molecules from the entire AIRCHECK DEL corpus enables better cross-target transfer learning compared to no pretraining approaches.
- We introduce three complementary OOD evaluation protocols (hierarchy, library, building-block splits) that systematically probe chemical novelty, cross-library transfer, and compositional generalization—key failure modes for practical DEL virtual screening.
- We show that self-supervised pretraining over fingerprint language leads to improved DEL binder prediction across six high-quality targets and quantify practical impact through enrichment analysis, achieving 13× enrichment at top-100.

2. Background

LMs for Chemistry. SMILES strings (Weininger, 1988) are textual representations of molecular structures that have enabled the application of NLP methods to accelerate drug discovery in various applications, such as molecular property prediction, drug-target interaction prediction, and hit generation (Yu et al., 2021; Wang et al., 2020; Hu et al., 2024; 2025; Ahmad et al., 2023). In particular, transformer-based models trained on SMILES, such as MOLBERT (Li & Jiang, 2021), ChemBERTa (Chithrananda et al., 2020), and MoLFormer (Wu et al., 2023), have made significant advances in the field of computational chemistry (Mswahili & Jeong, 2024). Further advances in data efficiency under limited labeled data settings can improve performance, particularly on complex tasks or rare molecular properties (Mswahili & Jeong, 2024). Pre-training of transformer-based models builds a deep understanding of relationships, patterns, and features within the data structures, enabling the model to perform well on broad range of in-distribution and out-of-distribution downstream tasks (Hendrycks et al., 2020).

Fingerprint Representations. Molecular fingerprints can be computed from SMILES strings, where each method aims to capture and encode different aspects of a molecule (Cereto-Massagué et al., 2015). Virtual screening (Rester, 2008) methods using molecular fingerprints leverage molecular similarity to known hits and non-hits. There are several types of molecular fingerprints, including substructure key-based fingerprints (Durant et al., 2002), path-based fingerprints (Dalke, 2019), and circular topological fingerprints (Morgan, 1965). Fingerprint representations are widely used in DEL hit discovery pipelines, with large-scale fingerprint datasets published on the AIRCHECK platform (Edwards & Owen, 2025). Importantly, the fingerprint-only setting introduces unique challenges: it precludes the use of existing molecular encoders that depend on structural information

(including graph neural networks and SMILES-based transformers), and existing fingerprint-based approaches rely on supervised tree-based models that cannot leverage unlabeled data for self-supervised learning.

DEL-ML Hit Generation. Virtual screening models, trained from DEL libraries, forming a DEL-ML pipeline (McCloskey et al., 2020; Iqbal et al., 2024; Han et al., 2024), can be used to screen even larger commercial libraries of small molecules, enabling the selection, purchase, and testing of hit compounds (Quigley et al., 2024). The DEL-ML pipeline avoids the need for resynthesis of original library hits and enables filtering of hit compounds based on chemical properties prior to selection and purchase (Wellnitz et al., 2024). With fingerprint DEL data from AIRCHECK, Wellnitz et al. (2024) show tree-based and boosting ensemble methods find success in a virtual screening task predicting hits for target WDR91. Leveraging fingerprints, alongside with SMILES, from BELKA (Quigley et al., 2024), Shlepov (2024) show that small transformer architectures can be fine-tuned to predict hits for a set of 3 targets. A key limitations remains in OOD prediction, where current models fail to effectively generalize to chemicals distinct from the training set (Blevins et al., 2024; Wellnitz et al., 2024).

3. Methods

We propose DELBERT-2, a foundation model for DEL virtual screening that learns representations from molecular fingerprints. DELBERT-2 is pretrained with self-supervised objectives on large collections of unlabeled molecules and then fine-tuned for downstream DEL hit prediction. The central idea is to convert sparse fingerprint vectors into a compact token sequence (a *fingerprint language*) suitable for transformer-based masked language modelling. See Appendix B for a detailed comparison of DELBERT vs DELBERT-2 methodology.

We pretrain on approximately 2.5M unlabeled molecules drawn from public DEL collections in the AIRCHECK database (Edwards & Owen, 2025). Unlike per-target pretraining approaches, this unified approach enables the model to learn generalizable molecular patterns across diverse protein targets and library designs, facilitating better cross-target transfer learning.

3.1. Chemical Fingerprints

Fingerprint Tokenization. We use molecular fingerprints rather than SMILES because AIRCHECK releases only pre-computed fingerprints to preserve confidentiality of proprietary structures, precluding SMILES-based encoders (Cereto-Massagué et al., 2015). Given pre-computed fingerprints (ECFP4/6, FCFP4/6, AtomPair, TopTor, and MACCS), we represent each active bit as a discrete token `fp_pos_val`, where `fp` is the fingerprint type, `pos` is the active bit index, and `val` is its value. This transforms high-dimensional sparse vectors into variable-length token sequences suitable for transformer-based masked language modelling.

ECFP4 and **ECFP6**, are extended-connectivity fingerprint with diameter 4 and 6, respectively. They are circular Morgan fingerprints that encode the local circular environment around each atom up to some number of bonds away (Morgan, 1965). In ECFP, atoms are labeled with detailed atomic invariants, such as element, valence, and connectivity, before being hashed into fingerprint bits. **FCFP4** and **FCFP6**, Functional-Class Fingerprint, are similar circular fingerprints to ECFP, but instead label atoms using functional classes, such as hydrogen-bond donor/acceptor, aromatics, and ionization, before being hashed to fingerprint bits. In general, ECFP is better suited for similarity searching and clustering, while FCFP is more appropriate for pharmacophoric applications.

AtomPair fingerprints are topological fingerprints that enumerate all pairs of atoms in a molecule, encoding the atom identifier at each end, and the topological distance, the number of bonds in the shortest path (Carhart et al., 1985). These atom-pair descriptions are then hashed into a fixed-length representation, capturing the global 2D shape of the molecule in terms of atom types and the distance between atoms. **TopTor** is another type of topological fingerprint, aimed at measuring the topological torsion in a molecule (Nilakantan et al., 1987). Each torsion is defined as the atom and bond type of four consecutively bonded atoms, before being hashed to fingerprint bits, thereby emphasizing the contiguous fragments in the molecule.

MACCS fingerprints use fixed dictionary-based structural keys to encode molecules based on the presence or absence of predefined substructures or properties (Durant et al., 2002). MACCS fingerprints are interpretable and are widely used in molecular similarity, clustering, and Quantitative Structure–Activity Relationship (QSAR) modelling (Cherkasov et al., 2014).

3.2. Architecture and Training

Like DELBERT (Seyed-Ahmadi et al., 2026), DELBERT-2 uses a BERT-style masked language modeling encoder; key improvements are: (i) *Unified pre-training:* 2.5M molecules across all AIRCHECK targets vs. 500K per-target; (ii) *Expanded fingerprints:* 8 types vs. 4; (iii) *Larger backbone:* ModernBERT (Warner et al., 2024) with 100M parameters (hidden 1024, 12 layers) vs. 70M (hidden 640, 16 layers); (iv) *OOD evaluation:* three protocols across 6 targets vs. library splits on 4.

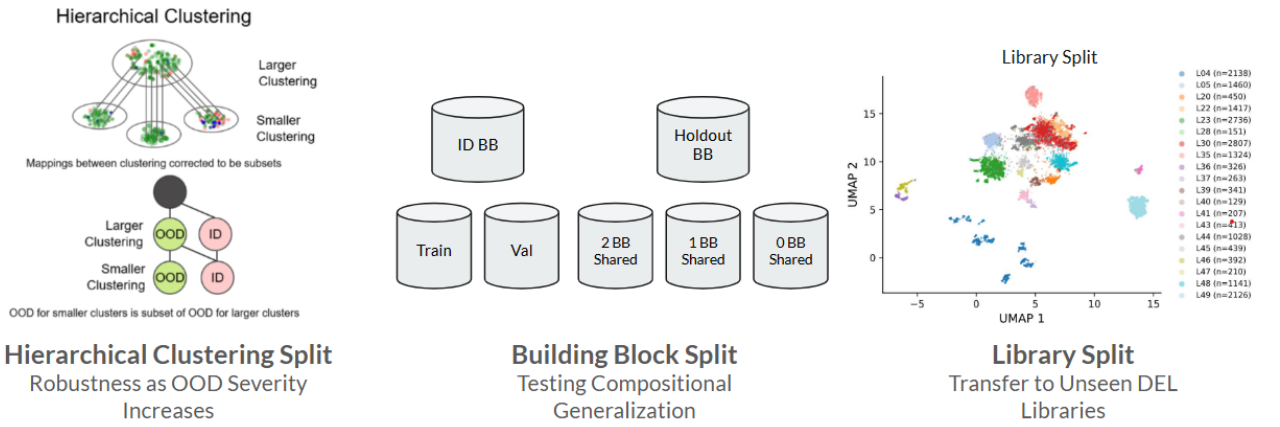


Figure 2. Out-of-distribution (OOD) evaluation protocols for DEL virtual screening. **Left: Hierarchical clustering split** constructs nested ID/OOD partitions at multiple Tanimoto similarity (Tanimoto, 1958) cutoffs (0.65/0.75/0.85); higher cutoffs create tighter clusters, making OOD test molecules more chemically distant from training. **Middle: Building-block split** holds out 40% of building blocks (BBs) and stratifies test molecules by BB overlap with training (`0bb`/`1bb`/`2bb`), directly probing compositional generalization. **Right: Library split** holds out entire DEL libraries, testing transfer to unseen synthesis designs (UMAP colored by library; n = number of molecules per library).

Model Architecture. DELBERT-2 is a 12-layer ModernBERT encoder (Warner et al., 2024) with hidden size 1024 ($\sim$100M parameters), followed by pooling and a binary classification head for fine-tuning. ModernBERT incorporates Flash Attention, rotary position embeddings (RoPE), and alternating local-global attention layers for efficient processing of long molecular sequences.

Self-supervised Pre training. Although AIRCHECK molecules carry binding activity labels, pre-training is entirely self-supervised on fingerprint structures. Fingerprint structures are used for masked language modelling (MLM, 15% masking, 20 epochs) on $\approx$2.5M molecules (Edwards & Owen, 2025). This unified strategy trains a single model on all available DEL data, enabling transferable representations across diverse targets and library designs.

Fine-tuning. For DEL hit prediction, we fine-tune using LoRA (Hu et al., 2021) (rank 8, α=16, applied to `Wqkv`), which freezes pre-trained weights and injects trainable low-rank matrices for parameter-efficient adaptation. We optimize with focal loss (γ=2) (Lin et al., 2017), which down-weights easy negatives to address the 5–11% positive class rate. Full hyperparameters are in Appendix **??**.

Ensemble LGBM baselines. We train LightGBM (LGBM) baselines on the same fingerprint features and train/validation/test splits used for DELBERT-2 to ensure fair comparison. We use an ensemble of K=5 LGBM models per run. Specifically, we apply stratified K-fold splitting to the training set and train one LGBM model on each fold's training partition, using the corresponding held-out partition for early stopping; final predictions are obtained by averaging predicted probabilities across ensemble members.

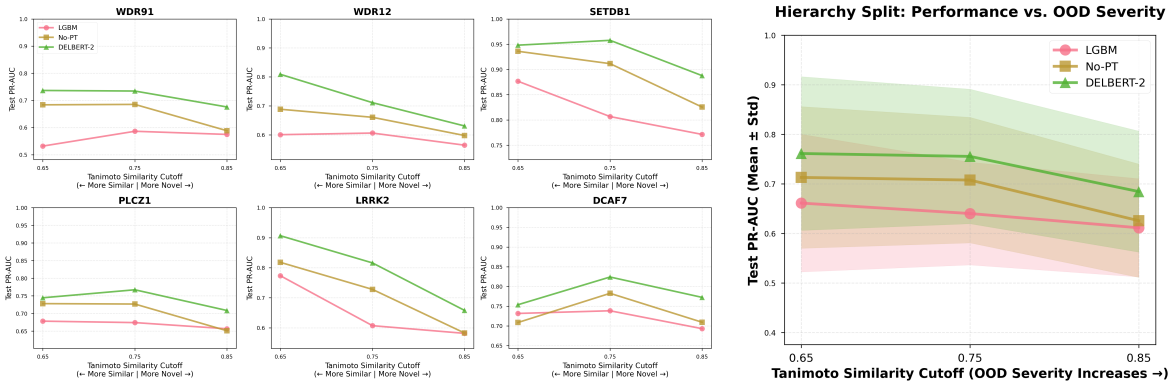
4. Experiments

We evaluate DELBERT-2 on DEL virtual screening with an emphasis on *out-of-distribution* (OOD) generalization. Our evaluation probes three failure modes: compositional generalization to unseen building blocks, transfer to unseen DEL libraries, and robustness to controllable chemical novelty. Unless stated otherwise, we compare (i) ensemble LightGBM (LGBM), (ii) the same transformer architecture without pre-training (No-PT), and (iii) DELBERT-2. Validation sets are in-distribution (ID); test sets are OOD by construction of each split protocol.

Datasets. We evaluate across six DEL fingerprint-only targets from AIRCHECK (WDR91, WDR12, SETDB1, PLCZ1, LRRK2, DCAF7). Each dataset contains on the order of 10^5–10^6 labeled molecules with substantial class imbalance, reflecting realistic DEL hit identification conditions. NSP was excluded from multi-panel figures due to consistently lower performance (PR-AUC $<$0.52 vs. $>$0.70 for other targets), though detailed results are included in supplementary materials.

4.1. Validation Protocols

Figure 2 summarizes the three complementary OOD protocols used in this work. As DEL chemistry is highly structured (e.g., through library design choices and building-block reuse), a single split strategy can substantially overestimate generalization if it preserves leakage pathways such as shared building blocks, shared libraries, or near-duplicate



(a) Per-target hierarchy split performance

(b) Mean across targets

Figure 3. Hierarchy split performance vs. OOD severity (Tanimoto similarity cutoff). **(a)** Per-target Test PR-AUC degrades as the cutoff increases (0.65 $\rightarrow$ 0.85), reflecting increasing chemical novelty. DELBERT-2 consistently outperforms LGBM and No-PT across all targets. **(b)** Mean $\pm$ std across 6 targets: DELBERT-2 maintains higher performance at all severity levels, with the largest advantage at the strictest cutoff (0.85), demonstrating that unified pre-training improves robustness to chemical novelty.

analogues. We therefore evaluate three complementary protocols that isolate different sources of distribution shift: (i) *hierarchy splits* probe robustness as chemical novelty increases; (ii) *library splits* test transfer across entire libraries with distinct synthesis designs; and (iii) *building-block splits* isolate compositional generalization when test molecules share fewer building blocks with training. Together, these protocols approximate realistic downstream deployment, where a model must generalize to new library chemistry, new building-block combinations, and increasingly novel chemical space.

Metrics for DEL Virtual Screening DEL hit prediction is both a classification task and a ranking task: in practice, a model is used to prioritize a small set of candidates for experimental follow-up. We therefore report PR-AUC as our primary threshold-free metric, since it is sensitive to false positives under extreme class imbalance. To evaluate early enrichment under a realistic validation budget, we also report NDCG@1000, which discounts binders found at lower ranks and emphasizes placing true binders near the top of the ranked list.

4.2. Results

We present results for each OOD protocol, highlighting discrimination (PR-AUC) and ranking quality (NDCG@1000). Across all protocols, DELBERT-2 generally improves over both LGBM and No-PT, with the clearest gains in stricter OOD settings.

4.2.1. Hierarchy Cluster Split: Robustness as OOD Severity Increases

The hierarchy split evaluates generalization under controllable distribution shift by clustering molecules in fingerprint space at different Tanimoto similarity cutoffs (0.65, 0.75, 0.85). Higher cutoffs typically correspond to more stringent notions of chemical novelty. We report hierarchy split PR-AUC and NDCG@1000 in Table 1.

Trends with increasing OOD severity. Across targets, performance generally degrades as the hierarchy split becomes more stringent (i.e., as test molecules become less similar to the training distribution), which is expected for DEL virtual screening under chemical novelty. Importantly, DELBERT-2's advantage over both LGBM and NPT is most pronounced in the stricter regimes, indicating that pretraining improves robustness rather than merely enhancing interpolation within the training distribution.

Practical implication. In real-world screening, the most valuable setting is precisely the strict OOD regime: the model must rank a small number of true binders near the top among a large candidate pool with limited overlap with the labeled DEL training data. The consistent gains in both PR-AUC and NDCG@1000 in Table 1 suggest that DELBERT-2 improves not only discrimination but also early enrichment, thereby directly translating into higher expected hit yield per experimental budget.

4.2.2. Library Split: Transfer to Unseen DEL Libraries

A library split is a cross-validation protocol where we hold out *entire DEL libraries* at test time. Each DEL dataset is a union of multiple libraries, and each library is produced by a specific combinatorial synthesis design (reaction scheme, building-block sets, and purification/selection characteristics). In a library split, all molecules from one or more libraries are assigned to the test fold, while training uses

Table 1. Hierarchy split results. **Left:** PR-AUC. **Right:** NDCG@1000. Higher is better. OOD severity increases with a higher clustering cutoff. DELBERT-2 consistently outperforms LGBM and NPT across targets, with the largest gains in stricter OOD regimes.

		PR-AUC					
		Cutoff 0.65		Cutoff 0.75		Cutoff 0.85	
Data	Model	ID	OOD	ID	OOD	ID	OOD
W91	LGBM	0.768	0.532	0.775	0.586	0.757	0.575
	NPT	0.827	0.684	0.828	0.685	0.784	0.588
	DBT-2	**0.881**	**0.737**	**0.878**	**0.735**	**0.865**	**0.676**
W12	LGBM	0.689	0.601	0.713	0.606	0.681	0.564
	NPT	0.769	0.688	0.784	0.661	0.714	0.598
	DBT-2	**0.812**	**0.809**	**0.821**	**0.711**	**0.749**	**0.631**
ST1	LGBM	0.898	0.877	0.896	0.807	0.902	0.771
	NPT	0.946	0.936	0.943	0.912	0.943	0.826
	DBT-2	**0.976**	**0.948**	**0.973**	**0.958**	**0.962**	**0.888**
PL1	LGBM	0.719	0.678	0.703	0.674	0.673	0.657
	NPT	0.746	0.728	0.733	0.727	0.682	0.651
	DBT-2	**0.786**	**0.745**	**0.767**	**0.767**	**0.709**	**0.709**
LK2	LGBM	0.719	0.773	0.718	0.607	0.691	0.582
	NPT	0.807	0.818	0.789	0.728	0.725	0.583
	DBT-2	**0.879**	**0.907**	**0.866**	**0.816**	**0.806**	**0.658**
DC7	LGBM	0.823	0.732	0.824	0.738	0.828	0.693
	NPT	0.837	0.709	0.835	0.783	0.826	0.709
	DBT-2	**0.851**	**0.754**	**0.852**	**0.824**	**0.852**	**0.772**

		NDCG@1000					
		Cutoff 0.65		Cutoff 0.75		Cutoff 0.85	
Data	Model	ID	OOD	ID	OOD	ID	OOD
W91	LGBM	0.943	0.832	0.941	0.779	0.832	1.000
	NPT	0.971	0.899	0.957	0.868	0.863	0.950
	DBT-2	**0.986**	**0.916**	**0.983**	**0.883**	**0.928**	0.997
W12	LGBM	0.750	0.896	0.785	0.680	0.878	0.778
	NPT	0.825	0.916	0.846	0.727	0.903	0.828
	DBT-2	**0.851**	**0.957**	**0.871**	**0.757**	**0.915**	**0.851**
ST1	LGBM	0.972	0.933	0.966	0.970	0.958	0.982
	NPT	0.985	0.971	0.984	0.987	0.969	0.935
	DBT-2	**0.999**	**0.984**	**0.994**	**0.994**	**0.983**	**0.987**
PL1	LGBM	0.920	0.838	0.895	0.841	0.831	0.967
	NPT	0.943	0.865	0.939	0.935	0.843	0.854
	DBT-2	**0.978**	**0.873**	**0.957**	**0.978**	**0.869**	**0.965**
LK2	LGBM	0.833	0.894	0.813	0.759	0.821	0.851
	NPT	0.894	0.915	0.875	0.860	0.845	0.774
	DBT-2	**0.956**	**0.966**	**0.937**	**0.941**	**0.895**	**0.938**
DC7	LGBM	0.944	0.894	0.939	0.860	0.903	0.899
	NPT	0.946	0.901	0.931	0.900	0.920	0.868
	DBT-2	**0.960**	**0.918**	**0.954**	**0.952**	**0.929**	**0.928**

Figure 4. Cross-library transfer performance. **Left:** Per-target breakdown showing DELBERT-2 consistently above LGBM and No-PT across exclusion fractions.**Right:** Mean ± std Test PR-AUC across 4 targets (WDR91, WDR12, SETDB1, LRRK2) as the held-out library fraction increases (0.1 → 0.3). DELBERT-2 maintains more consistent performance than baselines, demonstrating better cross-library transfer from unified pre-training.

molecules only from the remaining libraries.

Library splits provide a stronger and more realistic estimate of generalization than random or cluster-only splits because they reduce leakage from shared building blocks or closely related substructures across train and test sets. By holding out whole libraries, we directly test whether a model learns transferable structure-activity relationships that generalize beyond the idiosyncrasies of particular library designs.

Trends with increasing OOD severity. We report library split PR-AUC and NDCG@1000 in Table 2. As the held-out fraction increases (e.g., Split 0.10 → 0.30), the test set typically becomes more distribution-shifted relative to training and the task becomes harder. Across targets and split fractions, pretrained DELBERT-2 generally improves test PR-AUC and maintains or improves NDCG@1000, indicating better discrimination and stronger early enrichment when screening molecules drawn from unseen libraries.

Practical implication. In deployment, DEL-ML models are rarely used to re-rank close analogues from the same library chemistry used for training. Instead, they are used to prioritize candidates from *new* libraries or external catalogues whose synthesis designs and building-block compositions differ from the labeled DEL data available for a target program. Library-split performance therefore better reflects real-world utility: strong results indicate that the model has learned transferable structure–activity signals that generalize beyond library-specific biases.

Table 2. Library split results. **Left:** PR-AUC. **Right:** NDCG@1000. Split $s \in \{0.10, 0.20, 0.30\}$ denotes the approximate fraction of libraries held out. DELBERT-2 consistently improves over baselines, demonstrating robust cross-library transfer.

		PR-AUC					
		Split 0.10		Split 0.20		Split 0.30	
Data	Model	ID	OOD	ID	OOD	ID	OOD
LK2	LGBM	0.719	0.623	0.723	0.702	0.721	0.652
	NPT	0.807	0.565	0.810	0.754	0.814	0.614
	DBT-2	**0.874**	**0.721**	**0.855**	**0.776**	**0.856**	**0.745**
ST1	LGBM	0.896	0.889	0.896	0.893	0.894	0.872
	NPT	0.946	0.970	0.944	0.904	0.944	0.924
	DBT-2	**0.972**	0.887	**0.967**	**0.972**	**0.964**	**0.938**
W91	LGBM	0.779	0.557	0.773	0.639	0.769	0.580
	NPT	0.830	0.676	0.830	0.710	0.819	0.658
	DBT-2	**0.885**	0.650	**0.884**	**0.770**	**0.887**	**0.681**
W12	LGBM	0.704	0.626	0.661	0.565	0.644	0.620
	NPT	0.772	0.771	0.762	0.588	0.732	0.644
	DBT-2	**0.830**	0.675	**0.821**	**0.686**	**0.787**	**0.681**

		NDCG@1000					
		Split 0.10		Split 0.20		Split 0.30	
Data	Model	ID	OOD	ID	OOD	ID	OOD
LK2	LGBM	0.820	0.742	0.827	0.813	0.813	0.837
	NPT	0.900	0.716	0.899	0.815	0.905	0.773
	DBT-2	**0.957**	0.784	**0.939**	**0.908**	**0.928**	**0.944**
ST1	LGBM	0.970	0.971	0.966	0.980	0.966	0.982
	NPT	0.985	0.988	0.983	0.987	0.979	0.980
	DBT-2	**0.995**	**0.997**	**0.989**	**0.992**	**0.986**	**0.998**
W91	LGBM	0.946	0.646	0.936	0.812	0.915	0.818
	NPT	0.960	0.748	0.961	0.881	0.940	0.890
	DBT-2	**0.978**	**0.825**	**0.979**	**0.913**	**0.979**	**0.913**
W12	LGBM	0.780	0.750	0.771	0.612	0.792	0.792
	NPT	0.857	0.857	0.861	0.692	0.872	0.786
	DBT-2	**0.879**	**0.799**	**0.888**	**0.693**	**0.881**	**0.852**

Table 3. Building-block split results. 40% of building blocks are held out; test molecules are bucketed by overlap: **2BB** (both BBs seen), **1BB** (one seen), **0BB** (neither seen—strictest). Performance degrades as overlap decreases. DELBERT-2 often improves in stricter buckets, suggesting pre-training reduces building-block memorization. Relative: performance as % of ID validation.

		Validation			2BB Test			1BB Test			0BB Test		
Target	**Metric**	**LGBM**	**NPT**	**DBT-2**	**LGBM**	**NPT**	**DBT-2**	**LGBM**	**NPT**	**DBT-2**	**LGBM**	**NPT**	**DBT-2**
WDR91	PR-AUC	**0.875**	0.874	0.875	0.538	0.518	**0.542**	**0.298**	0.207	0.295	**0.174**	0.124	0.174
	Relative	100%	100%	100%	61%	59%	62%	34%	24%	34%	20%	14%	20%
WDR12	PR-AUC	0.731	**0.798**	0.735	**0.481**	0.350	0.461	**0.230**	0.159	0.211	**0.161**	0.145	0.142
	Relative	100%	100%	100%	66%	44%	63%	32%	20%	29%	22%	18%	19%
SETDB1	PR-AUC	0.943	**0.969**	0.932	0.604	**0.655**	0.601	0.293	**0.344**	0.309	0.218	0.187	**0.228**
	Relative	100%	100%	100%	64%	68%	64%	31%	36%	33%	23%	19%	**25%**
PLCZ1	PR-AUC	0.772	0.704	**0.776**	**0.706**	0.629	0.699	**0.624**	0.580	0.611	**0.547**	0.540	0.515
	Relative	100%	100%	100%	92%	89%	90%	81%	82%	79%	71%	77%	66%
LRRK2	PR-AUC	0.846	**0.878**	0.836	**0.636**	0.636	0.610	**0.418**	0.351	0.390	**0.223**	0.147	0.220
	Relative	100%	100%	100%	75%	72%	73%	49%	40%	47%	26%	17%	**26%**
DCAF7	PR-AUC	**0.865**	0.805	0.859	0.766	0.736	**0.767**	0.637	0.631	**0.646**	0.456	0.450	**0.488**
	Relative	100%	100%	100%	89%	92%	89%	74%	78%	75%	53%	56%	**57%**

4.2.3. Building-block Split: Compositional Generalization

DEL molecules are synthesized by joining building blocks (BBs) through a fixed reaction scheme. A building-block split is designed around how DELs are constructed. DEL molecules are synthesized combinatorially by joining a small number of building blocks through a fixed sequence of reactions. We hold out 40% of all BBs from training; test molecules are bucketed by overlap: `0bb` (no building blocks seen in training), `1bb` (one building block seen), and `2bb` (two building blocks seen). The `0bb` bucket tests compositional generalization to molecules assembled entirely from unseen building blocks. Note that AIRCHECK does not provide SMILES for individual BBs, only for assembled molecules, precluding BB-level similarity analysis.

Trends with increasing OOD severity. We report building-block split PR-AUC and NDCG@1000 in Table 3. As building-block overlap decreases (`2bb` → `1bb` → `0bb`), performance generally drops across all models, reflecting increasing difficulty. Importantly, DELBERT-2 improves PR-AUC and/or NDCG@1000 in the stricter buckets relative to LGBM and NPT, suggesting that pretraining helps reduce dependence on building-block memorization.

Practical implication. The `0bb` regime most closely matches the intended deployment setting: models must prioritize candidates from chemistry not well covered by labeled training data. At the same time, the consistently low absolute `0bb` scores across models highlight compositional generalization as a major bottleneck, motivating future work

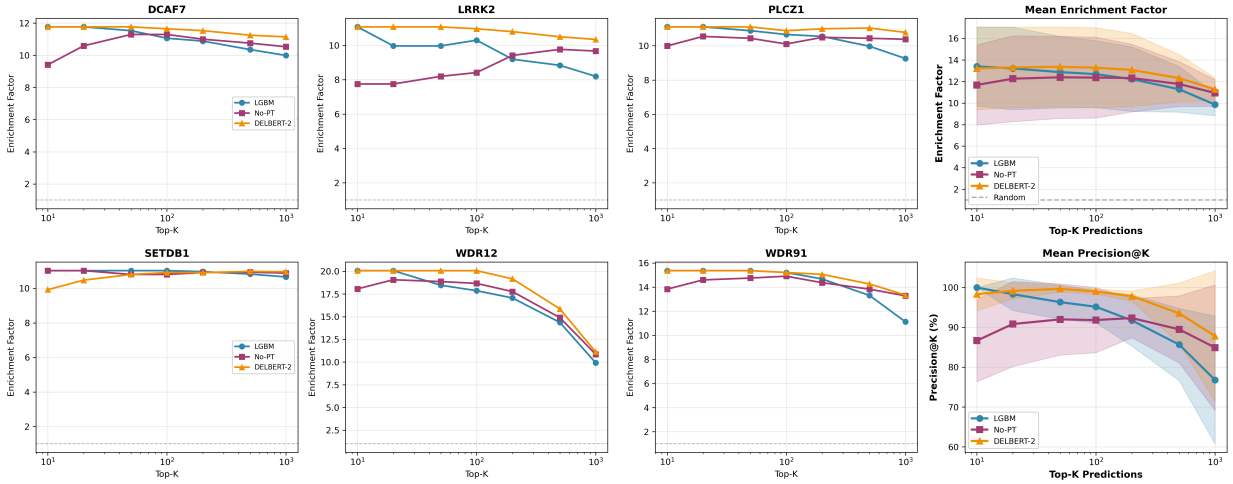


Figure 5. Enrichment analysis across targets. Top row: enrichment factors for DCAF7, LRRK2, PLCZ1, and mean across all 6 targets. Bottom row: SETDB1, WDR12, WDR91, and mean Precision@K. DELBERT-2 consistently achieves higher enrichment than baselines, with mean EF@100 of 13.28±3.72 vs. 12.36±3.75 (No-PT) and 12.69±3.11 (LGBM), representing 7.4% improvement over No-PT (p=0.040). All three models are evaluated at K ∈ {10, 20, 50, 100, 200, 500, 1000}.

on stronger pretraining and split-aware objectives.

4.2.4. Enrichment Analysis: Practical Screening Impact

To quantify the practical value of improved OOD performance for experimental validation, we analyze enrichment factors across all six targets and three OOD protocols. Enrichment measures how many more true binders are recovered in the top-K predictions compared to random selection, directly reflecting experimental screening efficiency.

Figure 5 shows enrichment curves across representative targets and OOD protocols. DELBERT-2 consistently outperforms baselines in early enrichment (top-100 to top-500), which is critical for practical screening where experimental validation budgets are limited. Across six targets, DELBERT-2 achieves 13.28±3.72 enrichment factor at top-100 compared to 12.36±3.75 for NPT (p=0.040, paired t-test), representing a 7.4% improvement. This translates directly to reduced screening costs: prioritizing the top 100 compounds predicted by DELBERT-2 captures approximately 13× more true binders than random selection, enabling substantial reduction in experimental validation burden while maintaining high recall.

Cross-protocol consistency. The enrichment advantage is maintained across hierarchy, library, and building-block splits, demonstrating that pretraining benefits are robust to different types of distribution shift. Even in the strictest 0BB regime where absolute performance is lowest, DELBERT-2 maintains higher enrichment than NPT, suggesting that the learned representations remain useful for hit prioritization even when compositional generalization is limited.

5. Discussion

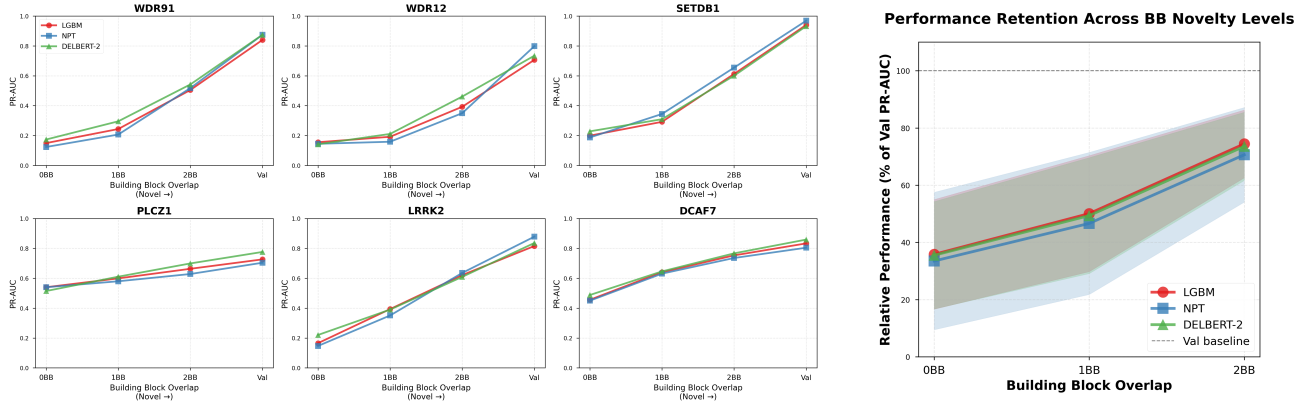
A key contribution of DELBERT-2 is the unified pretraining strategy: a single model is pretrained on 2.5M molecules spanning all available AIRCHECK DEL libraries, rather than training separate per-target models. This approach enables the model to learn generalizable molecular patterns that transfer across diverse protein targets and library designs, as evidenced by consistent improvements across six high-quality targets in Table 1.

The three validation protocols—hierarchy, library, and building-block splits—systematically probe different sources of distribution shift that arise in practical DEL virtual screening. Hierarchy splits measure robustness to chemical novelty; library splits test cross-library transfer; building-block splits assess compositional generalization. Together, these protocols provide a comprehensive characterization of OOD performance that cannot be captured by a single split strategy.

Across all three protocols, DELBERT-2 shows the largest gains over baselines in stricter OOD regimes (higher hierarchy thresholds, larger library exclusions, lower building-block overlap). This suggests that self-supervised pretraining learns transferable molecular representations that reduce reliance on memorizing specific training examples, building blocks, or library-specific patterns.

6. Limitations and Future Work

While DELBERT-2 demonstrates strong performance, several limitations remain. Compositional generalization remains challenging: absolute performance in the strictest



(a) Mean performance across targets

(b) Performance retention

Figure 6. Building-block compositional generalization analysis. **(a)** Test PR-AUC degrades as building-block overlap decreases (Val → 2BB → 1BB → 0BB), with DELBERT-2 showing better retention than baselines in stricter regimes. Lines show mean ± std across 6 targets. **(b)** Relative performance normalized to validation baseline. DELBERT-2 maintains higher retention in 0BB (20%) and 1BB (34%) buckets compared to NPT (14% and 24%), demonstrating that pretraining reduces building-block memorization.

building-block regime (0BB) is modest across all models, suggesting that pretraining alone is insufficient to fully address this failure mode.

Building-Block Compositional Generalization. Compositional generalization remains the primary limitation: absolute 0BB performance is modest across all models (PR-AUC 0.12–0.55), indicating that pre-training on assembled DEL molecules alone is insufficient to generalize to entirely unseen building blocks. As the model has no explicit knowledge of structural or pharmacophoric relationships between building blocks, it cannot transfer what it has learned about one BB to a chemically similar but unseen BB.

Multi-modal Architecture. Future work should explore multi-modal architectures that jointly model SMILES strings and fingerprints, which may capture complementary structural information. By combining SMILES tokenization with fingerprint language modeling, we can leverage both the interpretability of fingerprints and the rich sequential information in SMILES.

Explore Auxiliary Tasks. An auxiliary task predicting MACCS structural keys (167 binary features) provides explicit supervision on chemically meaningful substructures, potentially improving transfer to rare scaffolds. Preliminary experiments suggest multi-modal pretraining may reduce OOD degradation in building-block splits by learning more compositionally generalizable representations.

Expanded Pretraining Corpora. Current pretraining uses 2.5M DEL-specific molecules from AIRCHECK. Expanding to broader chemical databases (ChEMBL: 2.4M bioactive compounds, PubChem: 10M+ structures, MOSES: 1.9M drug-like molecules) could improve coverage of molecular scaffolds underrepresented in DEL libraries. This larger-scale pretraining (10M+ molecules) may particularly benefit generalization to novel chemical space in hierarchy and building-block splits.

Finally, prospective experimental validation in real-world screening campaigns with orthogonal binding assays and hit-to-lead optimization would establish whether computational OOD gains translate to improved experimental hit rates and reduced validation costs.

7. Conclusion

We introduced DELBERT-2, a pretrained fingerprint language model that converts sparse molecular fingerprints into unified token sequences and leverages ModernBERT self-supervised pretraining on 2.5M molecules from the entire AIRCHECK DEL corpus. Through unified pretraining, DELBERT-2 learns transferable molecular representations that generalize across six protein targets under three complementary out-of-distribution protocols: hierarchy splits probe chemical novelty, library splits test cross-library transfer, and building-block splits assess compositional generalization. Across these settings, DELBERT-2 consistently outperforms strong LightGBM ensemble baselines and comparable transformers trained from scratch, with particularly clear gains in stricter OOD regimes. These results demonstrate that fingerprint-centric self-supervised learning can effectively leverage unlabeled chemical structure data to improve hit prioritization under distribution shift, enabling practical DEL virtual screening deployment. More broadly, our findings highlight the need for comprehensive OOD evaluation for molecular models intended for real-world drug discovery applications.

References

Ahmad, S., Xu, J., Feng, J. A., Hutchinson, A., Zeng, H., Ghiabi, P., Dong, A., Centrella, P. A., Clark, M. A., Guié, M.-A., 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.

Blay, V., Tolani, B., Ho, S. P., and Arkin, M. R. High-throughput screening: today's biochemical and cell-based approaches. *Drug discovery today*, 25(10):1807–1821, 2020.

Blevins, A., Quigley, I. K., Halverson, B. J., Wilkinson, N., Levin, R. S., Pulapaka, A., Reade, W., and Howard, A. Neurips 2024 - predict new medicines with belka. https://kaggle.com/competitions/leash-BELKA, 2024. Kaggle.

Brenner, S. and Lerner, R. A. Encoded combinatorial chemistry. *Proceedings of the National Academy of Sciences*, 89(12):5381–5383, 1992.

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

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12):4977–5010, 2014.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Dalke, A. The chemfp project. *Journal of cheminformatics*, 11(1):76, 2019.

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

Edwards, A. M. and Owen, D. R. Protein–ligand data at scale to support machine learning. *Nature Reviews Chemistry*, pp. 1–12, 2025.

Gironda-Martínez, A., Donckele, E. J., Samain, F., and Neri, D. Dna-encoded chemical libraries: a comprehensive review with succesful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279, 2021.

Han, S., Guo, X., Wang, M., Liu, H., Song, Y., He, Y., Hsueh, K.-L., Cui, W., Su, W., Kuai, L., et al. Highly selective novel heme oxygenase-1 hits found by dna-encoded library machine learning beyond the del chemical space. *ACS Medicinal Chemistry Letters*, 15(9):1456–1466, 2024.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2744–2751, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.244. URL https://aclanthology.org/2020.acl-main.244/.

Hu, B., Saragadam, A., Layton, A., and Chen, H. Synthetic data from diffusion models improves drug discovery prediction. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 6278–6285. IEEE, 2024.

Hu, B., Layton, A., and Chen, H. Drug discovery smiles-to-pharmacokinetics diffusion models with deep molecular understanding. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

Hu, J. E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Iqbal, S., Jiang, W., Hansen, E., Aristotelous, T., Liu, S., Reidenbach, A., Raffier, C., Leed, A., Chen, C., Chung, L., et al. Del+ ml paradigm for actionable hit discovery–across del and cross ml model assessment. 2024.

Landrum, G. et al. Rdkit documentation. *Release*, 1(1-79): 4, 2013.

Li, J. and Jiang, X. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing*, 2021 (1):7181815, 2021.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

McCloskey, K., Sigel, E. A., Kearnes, S., Xue, L., Tian, X., Moccia, D., Gikunju, D., Bazzaz, S., Chan, B., Clark, M. A., et al. Machine learning on dna-encoded libraries: a new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866, 2020.

Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Mswahili, M. E. and Jeong, Y.-S. Transformer-based models for chemical smiles representation: A comprehensive literature review. *Heliyon*, 10(20), 2024.

Nilakantan, R., Bauman, N., Dixon, J. S., and Venkataraghavan, R. 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.

Quigley, I. K., Blevins, A., Halverson, B. J., and Wilkinson, N. Belka: The big encoded library for chemical assessment. In *NeurIPS 2024 Competition Track*, 2024.

Rester, U. From virtuality to reality-virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Current opinion in drug discovery & development*, 11(4):559–568, 2008.

Reza, S., Bagale, N., Wellnitz, J., Chiesa, L., Couñago, R. M., Melliou, S., Gordijo, C., Veenbaas, S., Chapman, J., Kaneva, V., et al. Aircheck: An open platform for ai-driven drug discovery. 2026.

Seyed-Ahmadi, A., Hu, B. X., Geraili, A., Layton, A., Chen, H. H., Kelley, S. O., and WANG, B. Delbert: Fingerprint language modeling for generalizable hit discovery in dna-encoded libraries. In *ICLR 2026 Workshop on Machine Learning for Genomics Explorations*, 2026.

Shlepov, V. 1st place solution [updated]. https://www.kaggle.com/competitions/leash-BELKA/writeups/victor-shlepov-1st-place-solution-updated, 2024. Kaggle competition "Predict New Medicines with BELKA" (BELKA).

Tanimoto, T. T. An elementary mathematical theory of classification and prediction. 1958.

Wang, J., Wang, H., Wang, X., and Chang, H. Predicting drug-target interactions via fm-dnn learning. *Current Bioinformatics*, 15(1):68–76, 2020.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., and Poli, I. 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.

Weininger, D. 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.

Wellnitz, J., Ahmad, S., Begale, N., Joseph, J., Zeng, H., Bolotokova, A., Dong, A., Reza, S., Ghiabi, P., Elisa, G., et al. Enabling open machine learning of dna encoded library selections to accelerate the discovery of small molecule protein binders. *Chemrxiv*, 2024.

Wu, F., Radev, D., and Li, S. Z. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5312–5320, 2023.

Yu, L., Su, Y., Liu, Y., and Zeng, X. Review of unsupervised pretraining strategies for molecules representation. *Briefings in functional genomics*, 20(5):323–332, 2021.

A. Fine-tuning Hyperparameters

Table 4 lists all fine-tuning hyperparameters used for DELBERT-2. LoRA (Low-Rank Adaptation) freezes the pre-trained model weights and injects trainable low-rank decomposition matrices into the attention layers, enabling efficient fine-tuning with far fewer parameters than full fine-tuning. Focal loss down-weights easy negatives and focuses training on hard-to-classify examples, which is critical given the 5–11% positive class rate in DEL datasets.

Table 4. Fine-tuning hyperparameters for DELBERT-2.

Hyperparameter	**Value**
LoRA rank	8
LoRA α	16
LoRA dropout	0.1
LoRA target module	`Wqkv` (attention)
Focal loss γ	2
Optimizer	AdamW
Early stopping metric	Validation PR-AUC
Positive class rate	5–11% (target-dependent)

B. DELBERT vs. DELBERT-2 Architecture Comparison

Table 5 summarizes the key architectural and training differences between the original DELBERT (Seyed-Ahmadi et al., 2026) and DELBERT-2. Both models use ModernBERT (Warner et al., 2024) as their backbone, with Flash Attention, RoPE, and alternating local-global attention. The key differences are model scale (DELBERT-2 is larger), fingerprint coverage (8 vs. 4 types), pre-training scope (unified 2.5M vs. per-target 500K), and fine-tuning strategy (focal loss + LoRA rank 8 vs. weighted BCE + LoRA rank 16).

Table 5. Architectural and training comparison between DELBERT (Seyed-Ahmadi et al., 2026) and DELBERT-2. Both use ModernBERT with Flash Attention and RoPE. Key improvements in DELBERT-2 include larger model capacity, expanded fingerprint coverage, unified pre-training on 5× more molecules, and comprehensive OOD evaluation.

Component	**DELBERT**	**DELBERT-2**
Backbone	ModernBERT	ModernBERT
Parameters	∼70M	∼100M
Hidden size	640	1024
Layers	16	12
Attention heads	16	16
Feed-forward size	1,152	1,152
Attention type	Alternating local-global (Flash Attention)	Alternating local-global (Flash Attention)
Position encoding	Rotary (RoPE)	Rotary (RoPE)
Fingerprint types	4 (ECFP4, FCFP6, AtomPair, TopTor)	7
Pre-training corpus	∼500K molecules (per-target, 100 epochs)	∼2.5M molecules (20 epochs)
Pre-training LR	5×10^{-4}	1×10^{-4}
Pre-training objective	MLM (15% masking)	MLM (15% masking)
LoRA rank / α	16 / 32	8 / 16
Loss function	Weighted BCE (pos_weight=15)	Focal loss (γ=2)
OOD protocols evaluated	Library splits	Hierarchy + Library + Building-block
Evaluation Target Datasets	4 targets	6 targets

Library Split Performance Comparison

Table 6 compares library split PR-AUC on the four overlapping targets. For DELBERT, we report the mean ± std across all K-fold CV folds (with per-target fold counts K in parentheses). For DELBERT-2, we report the OOD test PR-AUC at all

three library exclusion levels (Split 0.10, 0.20, 0.30), enabling a fuller picture of how DELBERT-2 performance varies with the fraction of libraries held out.

Table 6. Library split PR-AUC comparison between DELBERT (Seyed-Ahmadi et al., 2026) and DELBERT-2 on the four overlapping targets. DELBERT values are mean ± std across full K-fold library CV (all libraries cycled as test); K = number of CV folds per target. DELBERT-2 values are OOD test PR-AUC at three library exclusion levels (10%, 20%, 30%). **Note:** Baseline models differ: DELBERT uses XGBoost+LightGBM ensembles on binarized FPs (20 models/fold); DELBERT-2 uses LightGBM ensembles on count-based FPs (5 models/fold). DCAF7 is not available in DELBERT-2's library split evaluation.

	DELBERT OOD PR-AUC			**DELBERT-2 OOD PR-AUC**					
	full K-fold CV			Split 0.10		Split 0.20		Split 0.30	
Target	K	**Binary Baselines**	**DBT**	**LGBM**	**DBT-2**	**LGBM**	**DBT-2**	**LGBM**	**DBT-2**
WDR91	20	0.219±0.150	**0.349±0.162**	0.557	0.650	0.639	**0.770**	0.580	**0.681**
LRRK2	15	0.198±0.153	**0.336±0.215**	0.623	**0.721**	0.702	**0.776**	0.652	**0.745**
SETDB1	11	0.240±0.173	**0.585±0.250**	0.889	0.887	0.893	**0.972**	0.872	**0.938**
DCAF7	18	0.430±0.183	0.405±0.192	-	-	-	-	-	-

Differences in baseline performance. The most striking difference between the two papers is the absolute level of baseline performance. This large gap is primarily attributable to the evaluation protocol: DELBERT's full K-fold CV includes many folds where the held-out library is chemically very distant from all training libraries, producing low absolute scores. DELBERT-2's 20% holdout is less extreme on average. Additionally, DELBERT-2's baselines use count-based FPs (rather than binarized) and a larger expanded fingerprint set (8 types), which directly improves baseline performance.

Consistency of findings. Despite the protocol differences, both papers reach consistent qualitative conclusions: (i) self-supervised pretraining on fingerprint tokens substantially improves over tree-based baselines under library-based OOD evaluation; (ii) gains are largest on WDR91, LRRK2, and SETDB1, with DCAF7 being the exception where baselines are already competitive; and (iii) the improvement is most pronounced in early-enrichment metrics (EF@100, Prec@100, BEDROC in DELBERT; EF@100 in DELBERT-2), which are the most practically relevant for virtual screening.

C. Pre-training Configuration

Table 7 lists the full pre-training hyperparameters for DELBERT-2. Pre-training uses standard masked language modeling (MLM) with uniform random masking. Several optional features available in the codebase (span masking, segment-aware masking, MACCS auxiliary loss) were not used in the reported experiments.

Table 7. Pre-training hyperparameters for DELBERT-2.

Hyperparameter	**Value**
MLM masking probability	15%
Batch size (per GPU)	8
Gradient accumulation steps	8 (effective batch = 64)
Learning rate	1×10^{-4}
Optimizer	AdamW
Weight decay	0.01
Warmup ratio	10% of total steps
LR schedule	Cosine annealing (min ratio 0.1)
Gradient clipping	1.0
Precision	bf16-mixed
Pre-training epochs	20
Pre-training corpus	≈2.5M molecules

D. Hierarchy Split Algorithm

The hierarchical cluster split is the most technically novel evaluation protocol in this work. It proceeds in two phases: *cluster construction* (run once during data preparation) and *OOD sampling* (run at cross-validation time with different random seeds).

Cluster Construction. For each Tanimoto similarity cutoff $\tau \in \{0.85, 0.75, 0.65\}$, we apply the LeaderPicker algorithm (Landrum et al., 2013) to the ECFP4 fingerprints of all molecules in the dataset: This produces three independent cluster assignments at cutoffs 0.85 (coarse, $\sim$300 clusters), 0.75 (medium, $\sim$4K clusters), and 0.65 (fine, $\sim$30K clusters).

Hierarchy Rectification. Raw clusters at different cutoffs are not guaranteed to be nested. We apply a *rectification* step: for each fine-grained cluster, we find the majority-vote coarse cluster (the coarse cluster containing the most members of the fine cluster) and reassign all members accordingly. This ensures that fine clusters are strict subsets of coarse clusters, enabling hierarchical OOD propagation.

OOD Sampling. Given a random seed (different seeds produce different CV folds): Start with all top-level (coarse, τ=0.85) clusters. Sample 50% as OOD $\rightarrow$ these molecules form the test set for cutoff 0.85. **Propagate**: collect only the medium-level (τ=0.75) child clusters that belong to the OOD coarse clusters. Sample 25% of these as OOD $\rightarrow$ test set for cutoff 0.75. **Propagate again**: collect only the fine-level (τ=0.65) child clusters of the OOD medium clusters. Sample 10% as OOD $\rightarrow$ test set for cutoff 0.65. Non-OOD molecules are split 90/10 into train/val using stratified sampling.

Nesting Property. By construction, the OOD test sets are *nested*: the τ=0.85 test set $\subseteq$ the τ=0.75 test set $\subseteq$ the τ=0.65 test set. This enables controlled comparison across OOD severity levels without confounding by test set composition — the same molecules appear in all three test sets, with additional molecules added at looser cutoffs.

E. Building-block Split Details

The building-block split uses a *label-aware* holdout strategy to ensure the held-out building blocks include BBs that appear in positive (binding) molecules, preventing trivially easy splits where all positive molecules share training BBs.

Label-aware BB sampling. Of the 40% of BBs held out: 50% are sampled from BBs that appear in at least one positive molecule (binder) 50% are sampled randomly from the remaining BBs. This ensures the held-out set contains pharmacophoric BBs, making the 0BB test set genuinely challenging.

Train/test assignment. Each molecule is assigned a BB overlap score $k \in \{0, 1, 2, 3\}$ = number of its building blocks present in the training BB set: k=3: all BBs in training is validation; k=2: 2 BBs shared is medium OOD; k=1: 1 BB shared is high OOD; k=0: no BBs shared is strictest OOD.

Dataset limitation. AIRCHECK does not provide SMILES or structural information for individual building blocks; only for the assembled DEL molecules. This precludes BB-level similarity analysis (e.g., testing whether models generalize better to structurally similar held-out BBs). Future work with datasets that expose BB structures could enable finer-grained compositional generalization studies.

F. Library Split Details

Library prefix extraction. Each molecule's `LIBRARY_ID` encodes its library of origin (e.g., `L34_001` $\rightarrow$ library `L34`). We extract the library prefix by stripping the last underscore-delimited component, grouping all molecules from the same synthesis design.**Small library handling.** Libraries with fewer than 10 molecules are merged into a combined group that always remains in training. This prevents degenerate test sets from very small libraries while preserving the full dataset for training. **Exclusion percentages.** We evaluate at three exclusion fractions: 10%, 20%, and 30% of large libraries held out. Each exclusion level uses a deterministic seed derived from the base CV seed, ensuring reproducibility. No library appears in both train and test sets.