MGDEVAL: A CURATED MULTI-ENDPOINT BENCHMARK FOR GENERALIZATION IN MOLECULAR GLUE DEGRADER ACTIVITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular glue degraders (MGDs) are small molecules that induce or stabilize interactions between target proteins and E3 ubiquitin ligases, leading to target ubiquitination and subsequent proteasomal degradation. Their activity depends on the specific molecular and protein context, yet experimentally reported bioactivity remains fragmented across public resources, limiting systematic computational evaluation. We introduce MGDEval, an integrated data resource and benchmark dedicated to MGD activity prediction. Through manual curation, MGDEval integrates 19,815 records, including 9,893 with quantitative activity information, while preserving biological context and source evidence. To assess generalization to new scaffolds and target proteins, we benchmark 45 representation–model configurations for classification and regression across DC50, Dmax, EC50, and IC50. Classification performance declined with increasing distribution shift: mean AUROC fell from 0.895 under random splitting to 0.839 under scaffold splitting and 0.561 under target-cluster splitting. Analyses of this gap showed reduced chemical and protein support under target-cluster splitting and limited transfer gains from current protein representations, motivating models that learn how molecules, targets, and effectors jointly determine activity. Exact-label activity measurements and cell-line information improved generalization, highlighting the importance of data quality and biological context for prediction on unseen targets. MGDEval provides a reusable data foundation and reproducible evaluation framework for advancing activity prediction and guiding molecular glue discovery across new chemical and target spaces.

1 INTRODUCTION

Molecular glue degraders (MGDs) are small molecules that bring a target protein into contact with an E3 ubiquitin ligase. This interaction can trigger the destruction of the target protein, offering a way to remove proteins that are difficult to target with conventional drugs (Lu et al., 2026). Their activity depends on the molecule, its protein partners, and the experimental environment (Lu et al., 2024). MGD discovery has largely relied on serendipity and empirical screening. Computational models could guide discovery by predicting how active a molecule will be in a given biological setting. Yet only a handful of studies have addressed MGD activity prediction (Prael et al., 2024; Zhuang et al., 2026), compared with the broader research on PROTAC activity prediction (Ribes et al., 2026) and molecular glue structure modeling (Liao et al., 2025). These gaps highlight the need for a standardized MGD activity dataset and a benchmark aligned with the discovery goals that predictive models are intended to support.

Building such a dataset requires activity records with clear biological and experimental meaning. DC50 measures degradation potency, while Dmax measures the maximum extent of target removal. A molecule may act at low concentrations yet achieve limited degradation. EC50 and IC50 describe effects on protein function and cellular behavior, such as inhibition of cell proliferation. Together, the four endpoints provide complementary views of molecular glue activity. Existing databases contain these measurements, but they are scattered across entries, experimental descriptions, and patents. Some are reported as exact values, others as ranges, inequalities, or grades such as A, B, and C. Protein identities, cell lines, and assay conditions are often incomplete or ambiguous. Without resolving these differences and ambiguities, activity labels and biological inputs may not

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

be consistently defined for model training and comparison. Building a prediction dataset therefore requires recovering the meaning of each measurement and linking it to the correct molecule, proteins, and experimental setting (Landrum & Riniker, 2024).

A curated dataset provides the basis for model development, but evaluating the utility of predictive models for discovery also requires testing generalization beyond familiar compounds and proteins (Joeres et al., 2025). Two discovery goals motivate this evaluation. For established targets, researchers seek new molecular scaffolds as starting points for improving activity, selectivity, and drug-like properties. A more far-reaching goal is to find molecular glues for new targets, expanding the range of proteins and diseases accessible to this approach (Petzold et al., 2025). These goals require models to generalize to unseen chemical scaffolds and unseen target proteins, respectively. Existing studies have not systematically evaluated these capabilities, particularly transfer to unseen targets. Their performance on new scaffolds and targets therefore remains unclear.

We introduce MGDEval, an integrated data resource and benchmark dedicated to molecular glue degrader activity prediction. We integrate records from public databases and recover additional information from the underlying literature and patents. Through manual curation, we clarify activity measurements, resolve protein identities, and standardize experimental annotations. For modeling, the curated data are organized into 5,170 exact-label records, 4,723 interval-label records, and 9,922 activity-unlabeled records; the labeled records cover DC50, Dmax, EC50, and IC50.

To evaluate generalization across chemical scaffolds and target proteins, we establish classification and regression tasks across the four endpoints, together with a joint classification task for degradation potency and extent. We compare nine molecular–protein representation combinations with five predictive models under random, scaffold, and target-cluster splits. The target-cluster split holds out entire clusters of sequence-similar targets, testing whether models can transfer to target proteins beyond those represented in training. Further analyses examine data quality, experimental context, training-set similarity, and candidate ranking (Figure 1). Across four endpoints and 45 baseline configurations, mean classification AUROC decreases from 0.895 under random splitting to 0.839 under scaffold splitting and 0.561 under target-cluster splitting. Regression shows the same pattern: mean R^2 decreases from 0.646 to 0.499 and -0.190, respectively. These findings identify target generalization as a central challenge and establish MGDEval as a foundation for developing and evaluating models for new-target discovery.

2 Related Work

Molecular glue datasets and activity prediction. Recent databases, including MGDB, MGTbind, MolGlueDB, and TPDdb, have substantially expanded access to molecular glue knowledge and reported experimental evidence (Li et al., 2026; Zhu et al., 2026; Wang et al., 2026; Qin et al., 2026). These resources provide an important foundation for data-driven research, but do not provide shared benchmarks with unified activity tasks, evaluation splits, and reproducible baselines. Direct molecular glue activity prediction remains much less explored. Prael et al. (2024) showed that CRBN recruitment activity can be learned from large-scale screening data, demonstrating the feasibility of data-driven prediction across many targets. However, their evaluation used random splits within each target, leaving transfer to unseen targets insufficiently characterized.

Generalization in ternary-complex modeling. Recent computational studies have increasingly focused on ternary-complex structure modeling. DeepTernary (Xue et al., 2025) introduced an end-to-end model for PROTAC- and molecular glue-induced ternary structures, while Liao et al. (2025) systematically benchmarked cofolding methods on molecular glue complexes. Liao et al. further showed substantial memorization effects and significantly lower reconstruction performance on low-homology complexes, highlighting generalization as a broader challenge in molecular glue modeling. MGDEval brings this emphasis on generalization to activity prediction by providing a curated multi-endpoint dataset, standardized scaffold- and target-generalization settings, and reproducible baselines.

3 Data Curation and Resource Construction

Each MGDEval record is represented as $s = (M, T, E, C, X)$, where M, T, E, C, and X denote the molecule, target, effector, cell, and experimental context, respectively. Missing fields are retained explicitly. Activity endpoint and label are recorded separately when available.

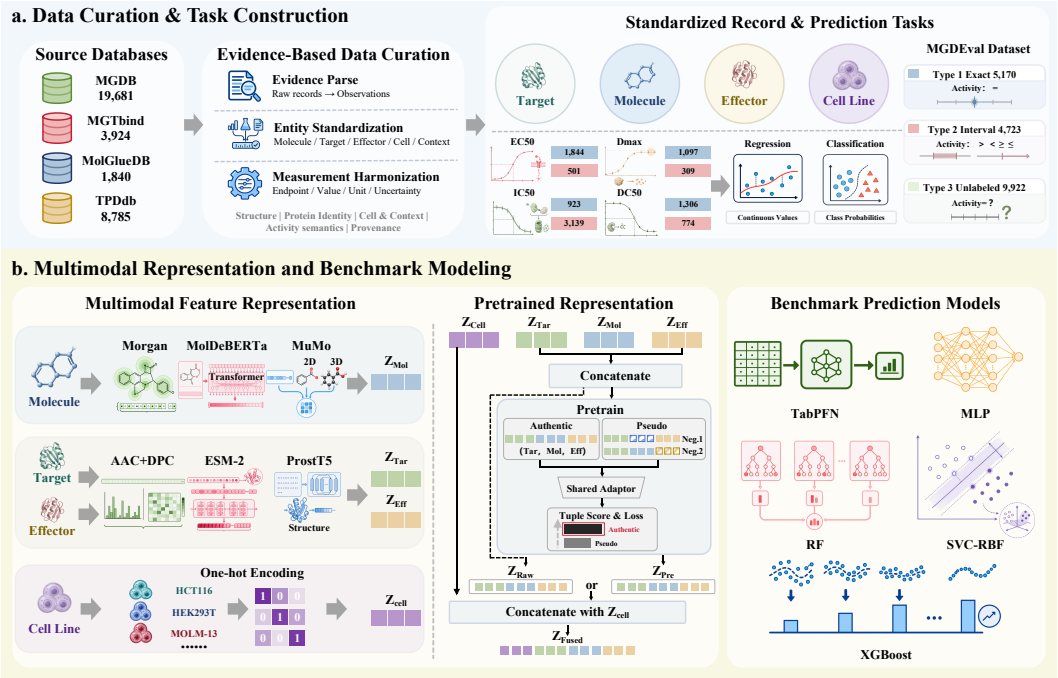


Figure 1: Overview of MGDEval data curation and benchmarking. (a) Manual curation preserves biological context and source evidence. Types 1–2 support activity prediction across four endpoints; complete unlabeled triplets from Type 3 support representation pretraining. (b) Molecular and protein representations, optionally transformed by a pretrained adaptor, are combined with cell-line encodings and evaluated using five predictive models.

We constructed MGDEval by tracing records from MGDB (Li et al., 2026), MGTbind (Zhu et al., 2026), MolGlueDB (Wang et al., 2026), and TPDdb (Qin et al., 2026) back to the corresponding literature and patents to recover measurement-level observations. This source-level reconstruction was necessary because the database records were not directly organized for a unified activity-prediction benchmark: a single entry may contain multiple proteins, cell lines, or activity values, while molecular identities, protein roles, assay conditions, or quantitative measurements may be incomplete or ambiguous. We mapped each recoverable measurement to its corresponding (M, T, E, C, X) and standardized molecular structures, protein identities, cell names, activity units, and comparison operators. Measurements from different experimental conditions were retained separately, whereas duplicate entries describing the same experiment were merged with their source links preserved.

Record types. We organize the curated records into three types according to the activity information associated with each record. Exact-label records report a numerical value for an activity endpoint; interval-label records report a range or bound rather than a single value; and activity-unlabeled records have no activity label but retain resolved molecule–target–effector identities. The three types contain 5,170, 4,723, and 9,922 records, respectively (Figure 1). The first two types provide 9,893 quantitative activity measurements for the supervised benchmark, while the unlabeled associations provide additional data for representation pretraining.

Resource overview. The 9,893 labeled activity records cover DC50, Dmax, EC50, and IC50 across 4,791 unique molecular structures. These records correspond to 8,881 unique (M, T, E, C, X) combinations. Endpoint overlap is sparse: 88.7% of these combinations have measurements for only one endpoint (Figure 2a), with DC50 and Dmax showing the largest cross-endpoint overlap. The five most frequent targets account for 59.9% of activity records, and the most frequent effector accounts for 80.1% (Figure 2b). Among labeled records, 6,944 contain cell-line information and 7,988 contain assay-context annotations. Context coverage differs between exact-label and interval-label records (68.14% vs. 94.54%), showing that label precision and annotation completeness capture distinct properties of the curated data. The 9,922 activity-unlabeled records correspond to 9,574

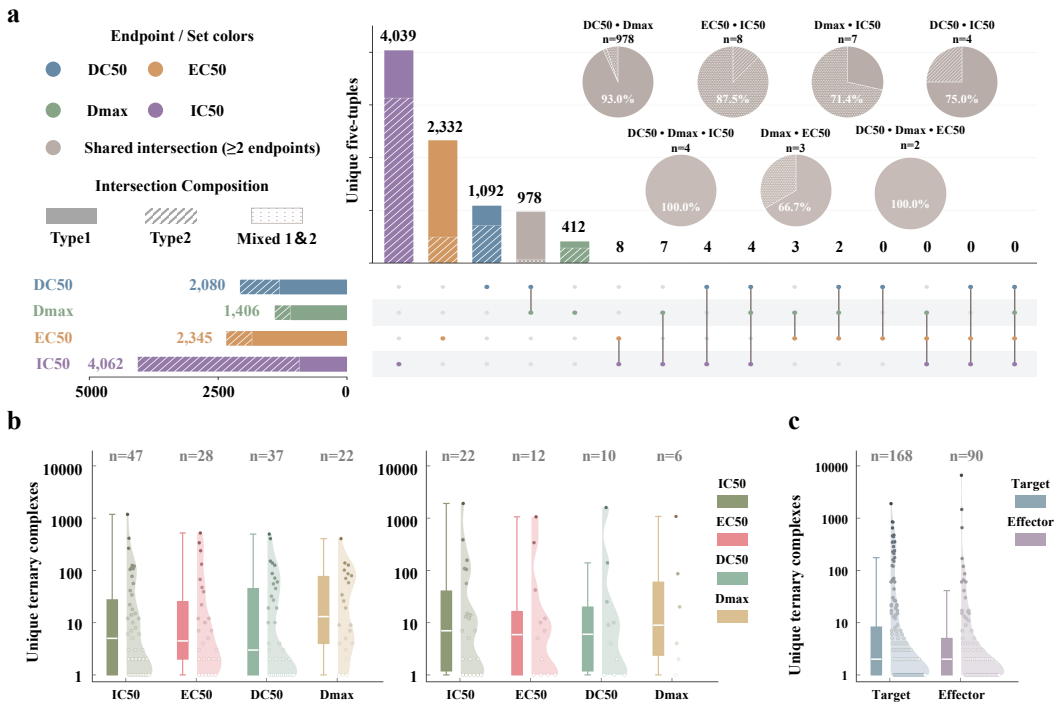


Figure 2: Endpoint overlap and protein coverage in MGDEval. (a) Endpoint overlap in Types 1–2, counted by unique molecule–target–effector–cell–context tuples. Horizontal and vertical bars show endpoint totals and exclusive intersection sizes, respectively; patterns and pies indicate type composition. (b) Distributions of unique molecule–target–effector triplet counts per target (left) and effector (right) in Types 1–2, grouped by endpoint. (c) Corresponding distributions for complete unlabeled Type 3 associations.

unique molecule–target–effector associations, providing additional association data for representation pretraining (Figure 2c).

4 MOLECULAR GLUE ACTIVITY PREDICTION BENCHMARKING

We benchmarked classification and regression across four activity endpoints under random, scaffold, and target-cluster splits to assess generalization to new chemical scaffolds and target proteins.

4.1 TASK CONSTRUCTION

We constructed classification and regression tasks for DC50, Dmax, EC50, and IC50 from Type 1 and Type 2 records. Each endpoint-specific sample was defined by a unique effector–molecule–target–cell line combination. Interval measurements were represented by their arithmetic midpoints and censored measurements by their reported numerical bounds; values sharing the same combination and endpoint were averaged. Binary classification used high-activity thresholds of DC50 $\leq$ 100 nM, Dmax $\geq$ 80%, EC50 $\leq$ 100 nM, and IC50 $\leq$ 100 nM. These thresholds, informed by activity criteria reported in the literature (Li et al., 2022) and patents (Dana-Farber Cancer Institute & Stanford University, 2025) and by the empirical distributions of the curated measurements, distinguish high-activity from not-high-activity samples. For regression, DC50, EC50, and IC50 were converted to p-scale values using $p = 9 - \log_{10}(c)$, where c is the numerical concentration in nM; Dmax was retained as a percentage. To identify candidates combining high degradation potency and extent, we also defined a joint classification task requiring both DC50 and Dmax criteria to be met (Table B3).

Table 1: Mean performance of the 45 baseline configurations across endpoints and evaluation splits.

	Random		Scaffold		Target-cluster	
Endpoint	AUROC ↑	R^2 ↑	AUROC ↑	R^2 ↑	AUROC ↑	R^2 ↑
DC50	0.854	0.523	0.797	0.376	0.551	−0.125
Dmax	0.873	0.555	0.800	0.399	0.524	−0.362
EC50	0.940	0.725	0.922	0.650	0.654	−0.044
IC50	0.913	0.780	0.837	0.571	0.514	−0.228
Overall	0.895	0.646	0.839	0.499	0.561	−0.190

Note: Values are means across the 45 baseline configurations, with each configuration averaged over cross-validation folds.

4.2 Evaluation protocol

Random evaluation used five-fold stratified cross-validation. Scaffold evaluation assigned molecules sharing a Murcko scaffold (Bemis & Murcko, 1996) to the same fold in a five-fold scheme. For target-cluster evaluation, proteins were clustered with MMseqs2 (Steinegger & Söding, 2017) and assigned by cluster to three folds. Fold assignment aimed to balance sample counts and class proportions as closely as possible while preserving group integrity, with activity distributions and group counts considered as secondary criteria.

We compared conventional descriptors with pretrained molecular and protein embeddings. Molecular representations comprised Morgan-1024 fingerprints (Rogers & Hahn, 2010), SMILES-pretrained embeddings (MolDeBERTa; de Oliveira & Saeed, 2026), and multimodal embeddings incorporating 2D/3D information (MuMo; Jing et al., 2025). Protein representations comprised amino acid and dipeptide composition (AAC+DPC), sequence embeddings (ESM-2; Lin et al., 2023), and embeddings from sequence–structure pretraining (ProstT5; Heinzinger et al., 2024). Target and effector proteins were encoded separately using the same representation, and model inputs were concatenated as $\mathbf{x} = [\mathbf{z}_{\text{effector}} \| \mathbf{z}_{\text{MGD}} \| \mathbf{z}_{\text{target}} \| \mathbf{c}_{\text{cell}}]$, where $\mathbf{c}_{\text{cell}}$ denotes cell-line one-hot encoding. The nine representation combinations were paired with MLP, random forest, RBF SVC/SVR (Pedregosa et al., 2011), TabPFN v3 (Grinsztajn et al., 2026), and XGBoost (Chen & Guestrin, 2016), yielding 45 configurations per task. Base representations were fixed, and all configurations shared predefined folds with preprocessing fitted on training data.

Classification performance was assessed using AUROC, average precision (AP), F1, recall, accuracy, and precision; regression performance was assessed using R^2, Spearman correlation, MAE, and RMSE. Performance was summarized as the mean across folds.

4.3 Benchmark results

Across four endpoints and 45 configurations, mean AUROC declined from 0.895 under random splitting to 0.839 under scaffold splitting and 0.561 under target-cluster splitting; corresponding mean R^2 values were 0.646, 0.499, and -0.190 (Table 1). IC50 had the highest mean random-split R^2 (0.780), but fell to -0.228 under target-cluster splitting. EC50 retained the highest mean target-cluster AUROC (0.654), despite a negative mean R^2 (-0.044).

No configuration was optimal across all four endpoints, and each endpoint's best configuration changed between random and target-cluster splits (Figure 3; Table 2). Among molecular representations, MolDeBERTa achieved the best marginal rank under random and target-cluster splits, whereas Morgan led under scaffold splitting but ranked last under target-cluster splitting. Morgan also had the highest mean R^2 under random and scaffold splits, but trailed both pretrained molecular representations on unseen targets. For protein representations, AAC+DPC led in marginal rank and mean R^2 across all splits, although ProstT5 achieved the highest mean target-cluster AUROC (Table D7).

Model rankings depended on the split and prediction task, with smaller differences in classification marginal ranks under target-cluster splitting (Table D7). For target-cluster EC50, SVC outperformed RF in mean AUROC across nine representation combinations (0.740 versus 0.671), but had lower F1 (0.261 versus 0.513). Thus, stronger discrimination did not ensure better high-activity classification at the evaluated decision threshold.

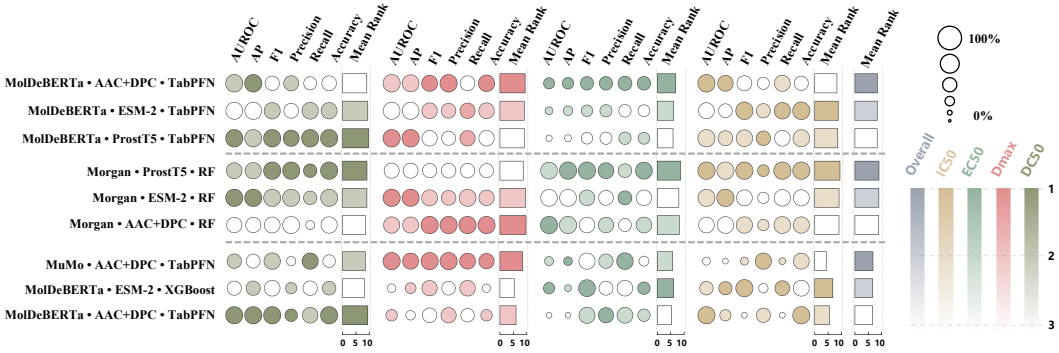


Figure 3: Top-three configurations by cross-endpoint aggregate ranking under the random (top), scaffold (middle), and target-cluster (bottom) splits across four endpoints and six evaluation metrics.

Table 2: Endpoint-wise best performance and corresponding configurations across evaluation splits.

		Random		Scaffold		Target-cluster	
Endpoint	Metric	Score	Configuration	Score	Configuration	Score	Configuration
DC50	AUROC↑	.886	MDB/PT5/TPFN	.823	MFP/ESM2/RF	.700	MDB/AD/TPFN
	R^2 ↑	.616	MDB/PT5/TPFN	.446	MDB/PT5/TPFN	.076	MDB/AD/RF
Dmax	AUROC↑	.899	MDB/PT5/TPFN	.824	MFP/ESM2/RF	.643	MFP/AD/TPFN
	R^2 ↑	.617	MFP/PT5/MLP	.476	MFP/AD/TPFN	−.125	MDB/AD/SVR
EC50	AUROC↑	.954	MMO/PT5/XGB	.937	MFP/AD/RF	.837	MFP/PT5/SVC
	R^2 ↑	.761	MMO/AD/TPFN	.682	MMO/PT5/TPFN	.197	MDB/PT5/TPFN
IC50	AUROC↑	.936	MDB/AD/TPFN	.865	MFP/PT5/RF	.642	MFP/PT5/SVC
	R^2 ↑	.820	MDB/AD/TPFN	.634	MFP/PT5/RF	−.056	MMO/PT5/RF

Note: Best fold-mean AUROC and R^2 among 45 baseline configurations, selected separately; configurations are shown as molecular/protein/model. AD: AAC+DPC; PT5: ProstT5; MDB: MolDeBERTa; TPFN: TabPFN. MMO: MuMo. MFP: Morgan-1024 fingerprint.

5 ANALYSIS OF UNSEEN-TARGET GENERALIZATION IN MGD ACTIVITY PREDICTION

5.1 ANALYSIS OF MODEL GENERALIZATION ACROSS CHEMICAL AND TARGET SPACES

To investigate the generalization gaps, we examined how splits altered chemical and protein support and how effectively predictors translated this support into transferable activity information.

Target-cluster splitting causes the largest drop in molecule similarity. We measured chemical support by the maximum Morgan–Tanimoto similarity to a training molecule. Among unique test molecules absent from training, mean similarity ranged from 0.746 to 0.765 across endpoints under scaffold splitting, compared with 0.399–0.504 under target-cluster splitting. Despite distinct scaffolds, 26.4% retained a training neighbor with similarity ≥ 0.8, while approximately 99% of scaffold-split test samples involved a previously seen target. Scaffold evaluation thus retained substantial chemical support around familiar proteins; holding out target clusters displaced both protein and chemical neighborhoods. EC50 illustrated how this retained support related to performance: its high-support fraction changed only from 45.0% under random splitting to 44.5% under scaffold splitting, accompanying the smallest AUROC decline among endpoints (0.018; Figures B3–B5).

EC50 performance depends on how many similar training molecules remain. Under target-cluster splitting, EC50 and IC50 had similar mean chemical support among novel test molecules (0.502 versus 0.504), but EC50 contained more low-support molecules and a larger high-support fraction (25.0% versus 3.6%). The mean therefore concealed two different distributions of training

support. Across EC50 folds, high-support fractions and AUROC declined together. An AUROC decomposition over all evaluated samples attributed 74.4% of EC50's net AUROC excess above 0.5 to positive–negative pairs in which both samples had high chemical support. Thus, its aggregate discrimination was concentrated in supported chemical neighborhoods despite substantial overall distribution shift.

Lower target similarity comes with larger performance drops. We quantified target support using the maximum number of identical residues in an MMseqs2 alignment to a training target, normalized by the length of the longer full sequence. Under target-cluster splitting, DC50 and Dmax had similarly sparse chemical support: mean similarity among novel test molecules was 0.423 and 0.399, respectively, and fewer than 0.4% had a training neighbor with similarity ≥ 0.8. Target support, however, differed substantially: 0.304 for DC50 versus 0.062 for Dmax. Dmax also showed larger random-to-target-cluster declines in AUROC (0.349 versus 0.303) and R^2 (0.917 versus 0.648), linking its greater transfer difficulty to a more distant protein domain.

EC50 provided a complementary comparison across splits. Among all test samples, its high-chemical-support fraction remained similar under scaffold and target-cluster splitting, while target support and AUROC declined sharply (Table 3A). Together, these comparisons suggest a joint chemical–protein domain of applicability: chemical neighborhoods provide local predictive support, while extending prediction to distant proteins requires learning how activity depends on molecular partners. Endpoint-level performance reflects where its observations lie within this joint domain.

Supervised models lose their advantage on unseen targets. We tested how much signal similarity alone recovered using one-nearest-neighbor prediction. Molecular retrieval was competitive mainly in target-cluster classification, reaching a mean AUROC of 0.571 against 0.561 for the supervised benchmark, whereas supervised models led under random splitting (0.895 versus 0.804; Table D4A). The key finding is the limited transfer gain from supervision: the advantage learned on familiar data did not translate into a consistent advantage over local retrieval on unseen targets.

Protein encodings show no clear advantage on unseen targets. To assess the contribution of protein features, we assigned each protein a fixed random embedding or shuffled embeddings among proteins. Across the evaluated predictors, the original protein encodings achieved slightly higher mean AUROC than random or shuffled embeddings under random and scaffold splitting, but showed no consistent advantage under target-cluster splitting. These results suggest that the information encoded by general-purpose protein representations does not automatically translate into transferable activity prediction. Progress may depend on learning how molecules, targets, and effectors jointly determine activity, beyond enriching the representations of individual proteins.

Association pretraining mainly helps the joint DC50–Dmax task. We therefore explored this relational information through adaptor pretraining, using the observed molecule–target–effector associations preserved in Type 3 despite lacking activity labels. Together with triplets pooled across labeled endpoints, these records supported label-free pretraining to rank observed combinations above sampled target or effector replacements while respecting each split's exclusion rules. The frozen adaptor output then served as the downstream representation. On the joint DC50–Dmax task, this increased mean target-cluster AUROC by 0.041 and AP by 0.032, while gains on other tasks were limited (Table D4). This benefit suggests that unlabeled associations and records from different endpoints contain shared molecular–protein relationship information that can support joint activity prediction.

Together, these analyses show that retained chemical and protein support helps explain generalization gaps across splits and endpoints. The limited cross-target gains from current representations, alongside the benefit of association pretraining on the joint task, motivate learning transferable molecular–protein relationships. Future models should capture how molecules, targets, and effectors jointly determine activity to extend prediction beyond familiar chemical and protein neighborhoods.

5.2 ANALYSIS OF DATA QUALITY AND CELLULAR CONTEXT

Beyond chemical and protein support, activity prediction depends on the measurements used for supervision and the experimental settings they describe. We next examined how evidence precision and cellular information shaped predictive performance.

Table 3: Training support, evidence quality, and cellular context in generalization.

Measure	Scope	Random	Scaffold	Target-cluster
A. Training support and performance				
High chemical support (%)	EC50	82.8	52.2	52.8
Mean target-sequence support	EC50	0.999	0.993	0.154
Mean AUROC	EC50	0.940	0.922	0.654
B. Regression gains (ΔR^2)				
Type 1 − Type 1+2 training	4-endpoint mean	+0.026	+0.037	**+0.183**
With − without cell-line features	4-endpoint mean	+0.018	+0.020	**+0.048**
With − without cell-line features	IC50	+0.024	+0.047	**+0.118**

Note: High chemical support denotes maximum training-set Tanimoto similarity ≥ 0.8; target support denotes full-length-normalized sequence matches.

Exact labels are critical for generalization on unseen targets. We compared Type 1-only training with Type 1+Type 2 training across the 45 benchmark configurations, evaluating both on the same Type 1 test samples. Averaged across the four endpoints, Type 1 training improved AUROC by 0.007 under both random and scaffold splitting, with a larger gain of 0.024 under target-cluster splitting. Regression gains were also largest under target-cluster splitting (Table 3B; Table D5). The choice of training evidence thus became particularly consequential when transferring to unseen targets.

EC50 showed the largest target-cluster classification gain: AUROC increased from 0.643 to 0.749, accompanied by an R^2 increase from -0.278 to 0.016. For IC50, however, Type 1 filtering reduced the sample pool from 3,957 to 700; R^2 improved while target-cluster AUROC declined, highlighting an endpoint-dependent trade-off between measurement precision and training coverage.

Cell-line information improves unseen-target generalization. We next assessed the contribution of the cellular setting by removing cell-line one-hot features and refitting the models on the same samples and folds. Cell-line information benefited EC50 and IC50 most, with EC50 showing the stronger classification gain. Averaged across all four endpoints, including cell identity improved classification and regression under all three splits, with the largest gains under target-cluster splitting. IC50 showed the largest endpoint-specific regression benefit (Table 3B; Table D5). Together with the Type 1 comparison, these results highlight the value of exact-label activity measurements and cellular information in the most challenging transfer setting.

The greater benefits of cell features for EC50 and IC50 are consistent with their roles in cellular assays, where they quantify downstream functional or inhibitory responses shaped by the selected cell line. The database's assay composition further helps explain EC50's stronger classification benefit: EC50 records are concentrated in cellular functional assays, whereas IC50 also encompasses cell-free biochemical or binding measurements. Among records lacking cell-line annotations, 44.2% of IC50 records explicitly belonged to binding or biochemical settings, compared with 2.6% of EC50 records. These findings motivate activity prediction that explicitly accounts for cellular and assay settings.

EC50 and IC50 show the largest cell-line gains. We next compared predictions for samples with reported cell lines against those without this annotation (C0). EC50 samples with cell annotations achieved higher AUROC under all three splits, with the gap increasing from 0.025 under random splitting to 0.072 under target-cluster splitting. This growing separation prompted us to examine how prediction errors differed between the two populations.

Within these populations, prediction errors also followed class composition. In EC50 and Dmax, samples with cell-line annotations were enriched for high activity, and their high-activity samples were more often correctly classified, whereas C0 samples were dominated by not-high-activity records and were more accurately classified within that class. Under scaffold splitting, mean recall for the high-activity EC50 class across the 45 configurations was 90.2% with cell-line annotations versus 34.3% without them (1,098 versus 77 high-activity samples). Thus, high overall accuracy in a population dominated by not-high-activity samples can conceal poor identification of high-activity candidates.

5.3 ANALYSIS OF CANDIDATE SELECTION AND PREDICTION CONFIDENCE

For experimental follow-up, a prediction's value depends on how it guides candidate selection. We assessed this utility through molecular ranking, joint degradation criteria, and confidence-based filtering.

Ranking performance for EC50 is largely retained on unseen targets. We compared activity rankings within groups sharing the same target, effector, and reported cell line, weighting groups equally. Under target-cluster splitting, EC50 predictions correctly ordered 64.7% of pairwise activity comparisons, compared with 49.9% for Dmax (Table D6A). EC50 thus retained information for relative prioritization despite weak regression performance, demonstrating the value of evaluating molecular ranking within a defined biological setting.

Low potency is harder to distinguish than insufficient degradation extent. Candidate selection requires both high potency (DC50 $\leq$ 100 nM) and sufficient maximal degradation (Dmax $\geq$ 80%). We evaluated the joint classifier by comparing compounds meeting both criteria with those meeting only one. Under scaffold splitting, AUROC was 0.840 against compounds with adequate potency but insufficient degradation extent. It was 0.630 against compounds with sufficient degradation extent but inadequate potency (Table D6B). This asymmetry held across all five classifier families, highlighting the need to verify potency when prioritizing compounds predicted to achieve sufficient degradation.

Confidence filtering is less effective on unseen targets. To assess whether confidence could guide candidate selection, we removed the highest-entropy 20% of predictions and measured error reduction. Relative to random splitting, target-cluster splitting increased predictive entropy in 86.9% of RF predictions and 88.3% of TabPFN predictions. Across the five classifiers, filtering reduced error by 5.3–6.0 percentage points under random splitting and 4.5–5.1 under scaffold splitting, but only 0.8–1.7 under target-cluster splitting. Thus, confidence-based filtering retained value on new scaffolds, whereas greater uncertainty on unfamiliar targets offered limited guidance for identifying individual errors. Temperature scaling (Guo et al., 2017), fitted on a training-side calibration set, reduced the fraction of all target-cluster MLP predictions that were both highly confident ($\geq$ 0.9) and incorrect from 24.9% to 9.0%. Calibration improved probability estimates while preserving within-classifier confidence ordering, highlighting the distinction between probability calibration and risk ranking. The practical value of uncertainty in guiding selection should thus be judged by error reduction at a fixed retention rate (Table D6C).

6 DISCUSSION AND CONCLUSION

We present MGDEval, an integrated data resource and benchmark dedicated to molecular glue degrader activity prediction. Evaluation across four complementary endpoints and multiple generalization settings identifies transfer to unseen targets as a central challenge.

Expanding molecular glue discovery requires both chemical diversification around established targets and the extension of degradation to new disease-relevant proteins. These goals make generalization a practical requirement for directing synthesis and biological testing. Ultimately, the value of activity prediction lies in the experimental decisions it improves: which compounds to advance, which activity properties to verify, and where additional evidence is most needed. By preserving biological context and source evidence, MGDEval provides a basis for understanding how molecular structure and protein partners jointly shape activity across experimental settings.

The current resource reflects uneven protein coverage, sparse endpoint overlap, and heterogeneous experimental conditions. Future progress will depend on expanding this evidence and learning how molecules, targets, and effectors jointly determine activity. Integrating these relationships with structural information and richer experimental context, alongside prospective validation, offers a path toward models that guide discovery across a broader range of proteins and biological settings.

AI USE STATEMENT

Generative AI tools were used to assist with coding and manuscript polishing. All scientific decisions, dataset curation, and source verification were performed or verified by the authors, who take full responsibility for the final work.

Reproducibility Statement

To facilitate reproducibility, we provide an anonymized repository containing the source code, benchmark partitions, configuration files, evaluation scripts, and instructions supporting the experiments reported in this paper: https://anonymous.4open.science/r/82078e9b-75b7-4dbe-870c-da6c97a34445/

References

Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* 1996, 39 (15), 2887–2893. https://doi.org/10.1021/jm9602928

Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785

Dana-Farber Cancer Institute, Inc.; Leland Stanford Junior University. CRBN molecular glue degraders and uses thereof. International Patent Publication WO2025129172A1, June 19, 2025. https://patents.google.com/patent/WO2025129172A1/en

de Oliveira, G. B.; Saeed, F. MolDeBERTa: Foundational model for physicochemical and Substructure-Informed molecular representation learning. *bioRxiv* 2026 (preprint; not peer reviewed). https://doi.org/10.64898/2026.02.15.706011

Ghosh, P.; Schmitz, M.; Pandurangan, T.; et al. Discovery and design of molecular glue enhancers of CDK12–DDB1 interactions for targeted degradation of cyclin K. *RSC Chemical Biology* 2025, 6 (1), 36–55. https://doi.org/10.1039/d4cb00190g

Grinsztajn, L.; Flöge, K.; Key, O.; et al. TabPFN-3: Technical report. *arXiv* 2026, arXiv:2605.13986 (version 2; not peer reviewed). https://doi.org/10.48550/arXiv.2605.13986 https://arxiv.org/abs/2605.13986v2

Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*; *Proceedings of Machine Learning Research*, Vol. 70; PMLR, 2017; pp 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

Heinzinger, M.; Weissenow, K.; Gomez Sanchez, J.; et al. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics* 2024, 6 (4), lqae150. https://doi.org/10.1093/nargab/lqae150

Jing, Z.; Sun, Y.; Li, Y. Y.; Janarthanan, S.; Deng, A.; Hu, P. Structure-aware fusion with progressive injection for multimodal molecular representation learning. In *Advances in Neural Information Processing Systems*, Vol. 38; 2025; pp 33247–33292. https://doi.org/10.52202/085713-0986

Joeres, R.; Blumenthal, D. B.; Kalinina, O. V. Data splitting to avoid information leakage with DataSAIL. *Nature Communications* 2025, 16 (1), 3337. https://doi.org/10.1038/s41467-025-58606-8

Kozicka, Z.; Suchyta, D. J.; Focht, V.; et al. Design principles for cyclin K molecular glue degraders. *Nature Chemical Biology* 2024, 20, 93–102. https://doi.org/10.1038/s41589-023-01409-z

Landrum, G. A.; Riniker, S. Combining IC_{50} or K_i values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling* 2024, 64 (5), 1560–1567. https://doi.org/10.1021/acs.jcim.4c00049

Landrum, G.; Tosco, P.; Kelley, B.; et al. rdkit/rdkit: 2025_09_6 (Q3 2025) release. RDKit: Open-source cheminformatics. Zenodo software release, 2026. Version 2025.09.6. https://doi.org/10.5281/zenodo.18797641

Li, C.; Luo, H.; Pang, Y.; et al. MGDB: A curated database for molecular glues. *Nucleic Acids Research* 2026, 54 (D1), D1488–D1499. https://doi.org/10.1093/nar/gkaf1131

Li, F.; Hu, Q.; Zhang, X.; et al. DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs. *Nature Communications* 2022, 13, 7133. https://doi.org/10.1038/s41467-022-34807-3

Liao, Y.; Zhu, J.; Xie, J.; Lai, L.; Pei, J. Benchmarking cofolding methods for molecular glue ternary structure prediction. *Journal of Chemical Information and Modeling* 2025, 65 (20), 11136–11148. https://doi.org/10.1021/acs.jcim.5c01860

Lin, Z.; Akin, H.; Rao, R.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023, 379 (6637), 1123–1130. https://doi.org/10.1126/science.ade2574

Lu, P.; Cheng, Y.; Xue, L.; et al. Selective degradation of multimeric proteins by TRIM21-based molecular glue and PROTAC degraders. *Cell* 2024, 187 (25), 7126–7142.e20. https://doi.org/10.1016/j.cell.2024.10.015

Lu, X.; Wang, X.; Yang, Z.; et al. Molecular glue degraders of HuR suppress BRAF-mutant colorectal cancer. *Nature* 2026, 655 (8123), 780–789. https://doi.org/10.1038/s41586-026-10613-5

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 2011, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Petzold, G.; Gainza, P.; Annunziato, S.; et al. Mining the CRBN target space redefines rules for molecular glue–induced neosubstrate recognition. *Science* 2025, 389 (6755), eadt6736. https://doi.org/10.1126/science.adt6736

Prael, F. J., III; Cox, J.; Sturm, N.; et al. Machine learning proteochemometric models for Cereblon glue activity predictions. *Artificial Intelligence in the Life Sciences* 2024, 6, 100100. https://doi.org/10.1016/j.ailsci.2024.100100

Qin, X.; Zhang, Y.; Wang, Y.; et al. TPDdb: The comprehensive database of targeted protein degrader. *Nucleic Acids Research* 2026, 54 (D1), D1683–D1691. https://doi.org/10.1093/nar/gkaf996

Ribes, S.; Dunlop, N.; Mercado, R. TACK: A statistical evaluation of degradation activity on a novel targeting chimeras knowledge dataset. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*; 2026; pp 11912–11923. https://doi.org/10.1145/3770855.3819052

Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* 2010, 50 (5), 742–754. https://doi.org/10.1021/ci100050t

Shanghai Meizer Pharmaceuticals Co., Ltd.; Fudan University. Compound having activity of degrading GSPT1 and application thereof. International Patent Publication WO2023078410A1, May 11, 2023. https://patents.google.com/patent/WO2023078410A1/en

Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* 2017, 35 (11), 1026–1028. https://doi.org/10.1038/nbt.3988

The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* 2025, 53 (D1), D609–D617. https://doi.org/10.1093/nar/gkae1010

Wang, X.; Zhuang, Z.; Zhang, C.; et al. MolGlueDB: An online database of molecular glues. *Nucleic Acids Research* 2026, 54 (D1), D1510–D1518. https://doi.org/10.1093/nar/gkaf811

Xu, Y.; Yu, Q.; Wang, P.; et al. A selective small-molecule c-Myc degrader potently regresses lethal c-Myc overexpressing tumors. *Advanced Science* 2022, 9 (8), 2104344. https://doi.org/10.1002/advs.202104344

Xue, F.; Zhang, M.; Li, S.; et al. SE(3)-equivariant ternary complex prediction towards target protein degradation. *Nature Communications* 2025, 16, 5514. https://doi.org/10.1038/s41467-025-61272-5

Zhu, J.; Liao, Y.; Lin, H.; et al. MGTbind: A comprehensive database of molecular glue ternary interactome. *Nucleic Acids Research* 2026, 54 (D1), D1500–D1509. https://doi.org/10.1093/nar/gkaf1075

Zhuang, Z.; Teng, D.; Xu, X.; et al. MG2Act: A mechanism-inspired sequential attention framework for molecular glue degradation prediction. *bioRxiv* 2026 (preprint; not peer reviewed). https://doi.org/10.64898/2026.08.08.739980

APPENDIX A. DATA CURATION AND DATASET OVERVIEW

A.1 DATA COLLECTION AND CURATION

We constructed MGDEval through evidence-based curation of molecular identities, protein roles, and experimental measurements. We traced ambiguous records to original articles, supplementary materials, and patents, examining chemical structures, assay methods, figure captions, and table headings to reconstruct measurement-level relationships. Compound identifiers and references from MGDB, MGTbind, MolGlueDB, and TPDdb supported source discovery and cross-referencing. Manual curation resolved ambiguous identities, protein roles, and assay–measurement assignments, supported by computational structure parsing, canonicalization, and unit conversion. Each curated measurement retained links to its source and supporting evidence.

Recovering molecular identity and measurement assignments. Structure curation matched compound identifiers and names to complete structures in the corresponding source material. We used these matches to resolve incomplete structural representations, then generated canonical SMILES preserving stereochemical information. This process recovered complete structures for 1,038 of 1,481 records examined for structural recovery. For entries containing multiple cells or activity values, we used local experimental descriptions and result ordering to establish cell–measurement assignments. Cell names such as RS4;11 and MV4;11 required interpreting semicolons as part of the name during record splitting. Examination of 743 observations yielded 64 corrected cell–measurement assignments. Cell normalization preserved source-reported genotypes and sorted subpopulations alongside the base cell identity.

Resolving protein roles and mixed measurements: SR-4835. One entry reported DC_{50} as `14_90` and the target as `CDK12_CCNK`, combining protein names and results from distinct experiments. We inspected the original assay methods and reporter constructs, identified cyclin K degradation as the measured response, and assigned CCNK as the target, DDB1 as the effector, and CDK12 as an auxiliary component. We assigned reference sequences using protein names, species, and accessions (The UniProt Consortium, 2025) and checked sequence validity and consistency. Reporter constructs and auxiliary proteins were recorded with the experimental conditions. The 14 nM result mapped to a HEK293T-Cas9 cyclin K–eGFP reporter assay measured by flow cytometry after 5 h. The approximately 90 nM result mapped to an A549 assay with endogenous N-terminal HiBiT-tagged cyclin K after 2 h; this experiment also reported $D_{max} > 95\%$. We reconstructed three measurements across two conditions, retained the approximation flag for DC_{50}, and represented the D_{max} constraint as $(95, 100]\%$. Matching the original result locations also identified duplicate source entries, whose provenance was consolidated (Kozicka et al., 2024; Ghosh et al., 2025).

Recovering quantitative evidence from an activity flag: WBC100. The source entry contained `IsActive=1`. We examined the original cell-viability experiments and Supplementary Figure S1, matched each value to its cell population, and checked protein identities against the reported mechanism and recombinant human CHIP experiments. This established human MYC as the target and CHIP/STUB1 as the effector. The 72 h experiments provided six point values, including 61 nM for Mia Paca-2 and 2,205 nM for L02; normal peripheral blood cells contributed a lower-bound constraint of > 150 nM. The resulting seven records paired six point values and one constraint with their cell identities, reported assay conditions, and individual evidence locations (Xu et al., 2022).

Separating cell identity, dose, and endpoint: JeKo-1. The SCHEMBL25175866 entry contained "Jeko-1(100 nM)" and "Jeko-1(10 μM)" as cell descriptions, associated with values of 35% and 53%. We located the compound in Table 1 of WO2023078410A1 and examined Test Example 1 and the table headings. The evidence identified an immunoblot assay following 24 h treatment and assigned the two percentages to degradation at specified doses. We interpreted both descriptions as the same JeKo-1 cell identity, extracted doses of 100 and 10,000 nM, and paired each percentage with its dose. The endpoint was recorded as fixed-dose degradation (Shanghai Meizer Pharmaceuticals & Fudan University, 2023).

Interpreting target-specific patent grades. Table 1 of WO2025129172A1 reports both CK1α and WEE1 measurements, requiring target-specific identification of the relevant column. We located

compounds 5 and 6 in the WEE1 DC_{50} column and read the associated grading legend, mapping "+++" to < 100 nM and "+" to 1,000–10,000 nM. Examination of Example 2 established the Jurkat HiBiT reporter system and a 5 h compound exposure; the separately reported 10 min referred to detection-reagent incubation. We paired each quantitative constraint with its target, reporter system, and treatment duration, retaining unspecified boundary inclusivity where the patent left interval endpoints undefined (Dana-Farber Cancer Institute & Stanford University, 2025).

When activity information was embedded in assay descriptions or contextual text, we extracted the endpoint, numerical value, unit, and comparison operator and matched them to the corresponding compound and experimental condition. We used the surrounding assay description to distinguish measured activity from treatment doses and reagent concentrations. Across measurements, concentrations were standardized to nM and degradation levels to percentages, retaining original values, approximation flags, and interval boundaries. Records were classified by quantitative evidence and the completeness of molecular and protein information, with separate categories for unlabeled associations, alternative measurements, and unresolved records.

Consolidating comparable observations. We grouped measurements by endpoint and the combination of molecule, target, effector, cell, and experimental context. Original figures, compound identifiers, and assay descriptions supported source deduplication, while measurements under distinct conditions retained separate representations. Representative selection considered point values, interval informativeness, and directly comparable activity. We compared 404 measurements across 130 groups, selected a representative for each group, and preserved alternative measurements with their provenance and selection rationale. Source links, curation decisions, selection rules, field definitions, and complete mappings are included in the MGDEval release package.

A.2 Dataset Composition and Experimental Context

In addition to Types 1–3, MGDEval retains 17,382 additional curated records, summarized in Table A1.

Table A1: Additional curated records retained outside Types 1–3. Percentages are calculated over the additional 17,382 records shown.

Category	Definition	Records	Share (%)
Target missing or unresolved	Target identity is missing or cannot be resolved	6,949	39.98
Effector missing or unresolved	Effector identity is missing or cannot be resolved	3,507	20.18
Target and effector missing or unresolved	Both target and effector identities are missing or unresolved	6,023	34.65
Other or unresolved endpoint	Records outside the four benchmark endpoints or with unresolved endpoint assignment	903	5.20
Total		**17,382**	**100.00**

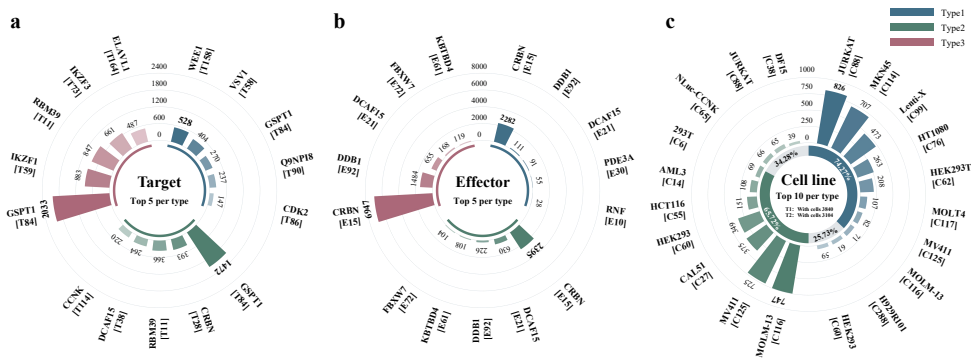


Figure A1: Distribution of frequently represented entities across data types. Record counts for the five most frequent targets (a) and effectors (b) in each of Types 1–3, and the ten most frequent cell lines (c) in Types 1–2. Inner rings in (c) indicate the proportions of records with and without cell-line annotations.

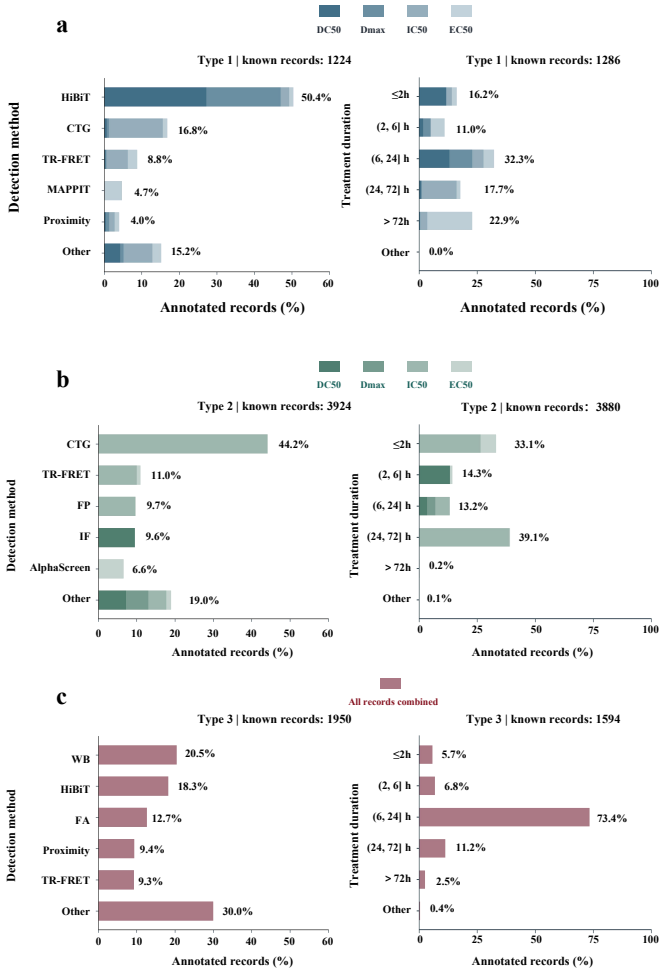


Figure A2: Assay detection methods and treatment durations across data types. Distributions of detection methods (left) and treatment durations (right) in Type 1 (a), Type 2 (b), and Type 3 (c). Percentages are calculated separately within each panel using records with the corresponding annotation; panel titles report the denominators. Stacked segments in Types 1–2 indicate contributions from DC50, Dmax, IC50, and EC50 records.

B TRAINING DATA AND CROSS-VALIDATION

Sample construction. For each endpoint, benchmark samples were defined by unique molecule–target–effector–cell line combinations. Missing cell line was treated as a separate category. Point measurements retained their numerical values; intervals were represented by their midpoints and inequalities by their numerical bounds. Averaging was performed in the original units, with equal weight per measurement. This yielded 9,398 endpoint-specific samples from 9,893 measurements, spanning 8,155 distinct quadruples (Table B1). Positive labels denote DC50/IC50/EC50 ≤ 100 nM or Dmax $\geq 80\%$. Groups at the cutoff were labeled negative when all constituent measurements specified concentration > 100 nM or Dmax $< 80\%$.

Table B1: **Supervised training cohorts.** Record counts are reported before aggregation; samples are endpoint-specific quadruples. Positive entries give count (percentage).

	Source records		Aggregated samples			Groups	
Endpoint	Type 1	Type 2	Total	Positive (%)	Negative	Target	Scaffold
DC50	1,306	774	1,959	1,076 (54.9)	883	32	801
Dmax	1,097	309	1,319	685 (51.9)	634	19	538
EC50	1,844	501	2,163	1,175 (54.3)	988	25	545
IC50	923	3,139	3,957	1,206 (30.5)	2,751	43	1,313

Cross-validation. Each endpoint was evaluated with stratified random 5-fold (seed 42), Murcko scaffold 5-fold, and target-cluster 3-fold partitions. Scaffolds retained atom and bond types with stereochemistry omitted; empty scaffolds formed one group. Targets were clustered separately for each endpoint using MMseqs2 (sequence identity 0.5, coverage 0.8, coverage mode 0, cluster mode 0, sensitivity 7.5, and cluster reassignment). Scaffold and target groups were assigned intact to nonempty folds by minimizing

$$\mathcal{L}_{\text{split}} = \sum_{f=1}^{K} \sum_{m} w_m \frac{|n_{mf} - N_m/K|}{\max(N_m/K, 1)}, \tag{B1}$$

where n_{mf} and N_m are fold-specific and total counts. Weights were 4 for sample count, 2 each for positive and negative counts, 0.35 for each of five activity bins, and 1 for group count. Activity bins followed quintile cut points, using $\log_{10}(\max(a, 10^{-12}))$ for concentration endpoints and raw Dmax values. Each fold served once as validation, with the remaining folds used for training (Table B2).

Table B2: **Cross-validation fold composition.** Entries report validation sample count (positive-class percentage). Each held-out fold is paired with the remaining samples for training.

Endpoint	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Random: stratified 5-fold					
DC50	392 (55.1)	392 (54.8)	392 (54.8)	392 (54.8)	391 (55.0)
Dmax	264 (51.9)	264 (51.9)	264 (51.9)	264 (51.9)	263 (52.1)
EC50	433 (54.3)	433 (54.3)	433 (54.3)	432 (54.4)	432 (54.4)
IC50	792 (30.4)	792 (30.6)	791 (30.5)	791 (30.5)	791 (30.5)
Scaffold: Murcko 5-fold					
DC50	392 (54.8)	392 (54.8)	392 (55.1)	392 (54.8)	391 (55.0)
Dmax	264 (51.9)	264 (51.9)	264 (51.9)	264 (51.9)	263 (52.1)
EC50	433 (54.3)	433 (54.3)	433 (54.3)	432 (54.4)	432 (54.4)
IC50	792 (30.4)	792 (30.6)	791 (30.5)	791 (30.5)	791 (30.5)
Target: cluster 3-fold					
DC50	656 (50.3)	652 (55.2)	651 (59.3)	–	–
Dmax	471 (71.3)	434 (41.5)	414 (40.8)	–	–
EC50	989 (79.7)	613 (27.9)	561 (38.5)	–	–
IC50	2,117 (41.6)	996 (15.3)	844 (20.5)	–	–

Table B3: Cohort and cross-validation fold composition of the joint DC50–Dmax classification task. The cohort contains 905 samples: 501 positive (55.4%) and 404 negative.

Split	Fold	Training	Validation	Positive (%)	Negative
Random	1	724	181	101 (55.8)	80
	2	724	181	100 (55.2)	81
	3	724	181	100 (55.2)	81
	4	724	181	100 (55.2)	81
	5	724	181	100 (55.2)	81
Scaffold	1	724	181	100 (55.2)	81
	2	724	181	101 (55.8)	80
	3	724	181	100 (55.2)	81
	4	724	181	100 (55.2)	81
	5	724	181	100 (55.2)	81
Target-cluster	1	449	456	305 (66.9)	151
	2	719	186	66 (35.5)	120
	3	642	263	130 (49.4)	133

Note: Positive and negative counts refer to the validation set. Each held-out fold is paired with all remaining samples for training.

Partition characteristics. All training and validation sets contained both classes. Quadruples were disjoint between training and validation, with scaffold and target-cluster disjointness enforced in the corresponding cold partitions. Random and scaffold partitions had similar class proportions across folds, with fold sizes differing by at most one sample. Target-fold imbalance reflects large indivisible clusters: the largest clusters accounted for 53.1% of IC50 and 43.7% of EC50 samples.

Joint DC50–Dmax classification. The joint task included unique molecule–target–effector–cell line combinations with both DC50 and Dmax labels available. Samples were positive when both high-activity criteria were met and negative otherwise. Cohort statistics and fold composition are reported in Table B3.

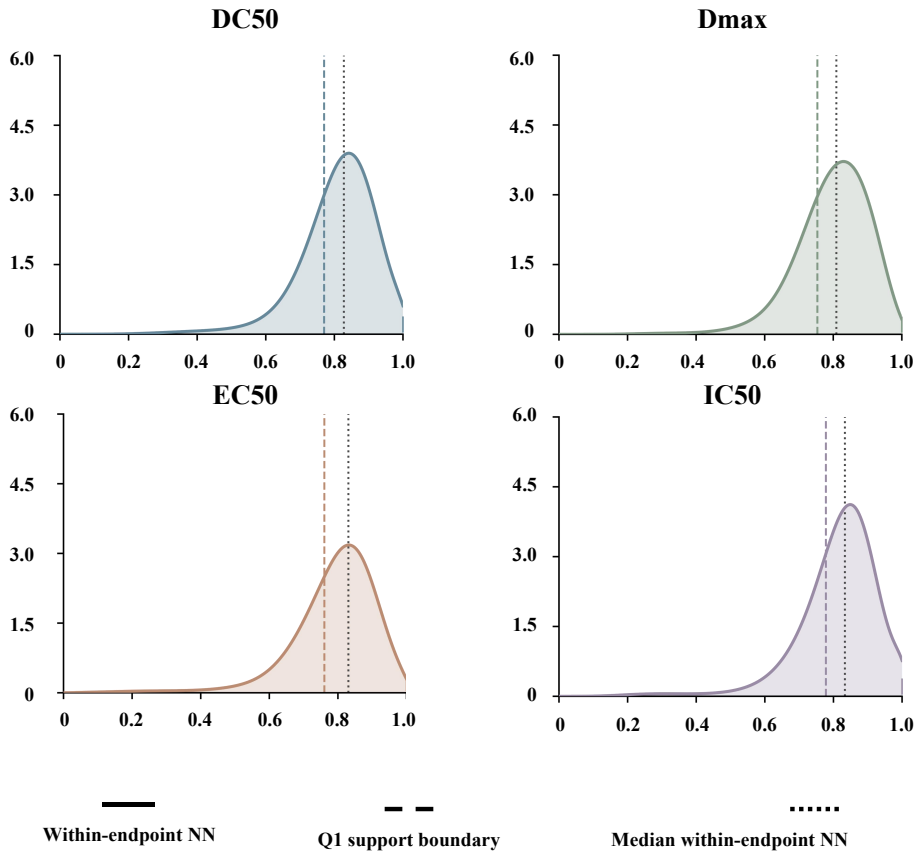


Figure B1: Within-endpoint chemical structural support. Distributions of nearest-neighbor molecular similarity within DC50, Dmax, EC50, and IC50. Dashed and dotted vertical lines indicate the first quartile (Q1 support boundary) and median of each distribution, respectively.

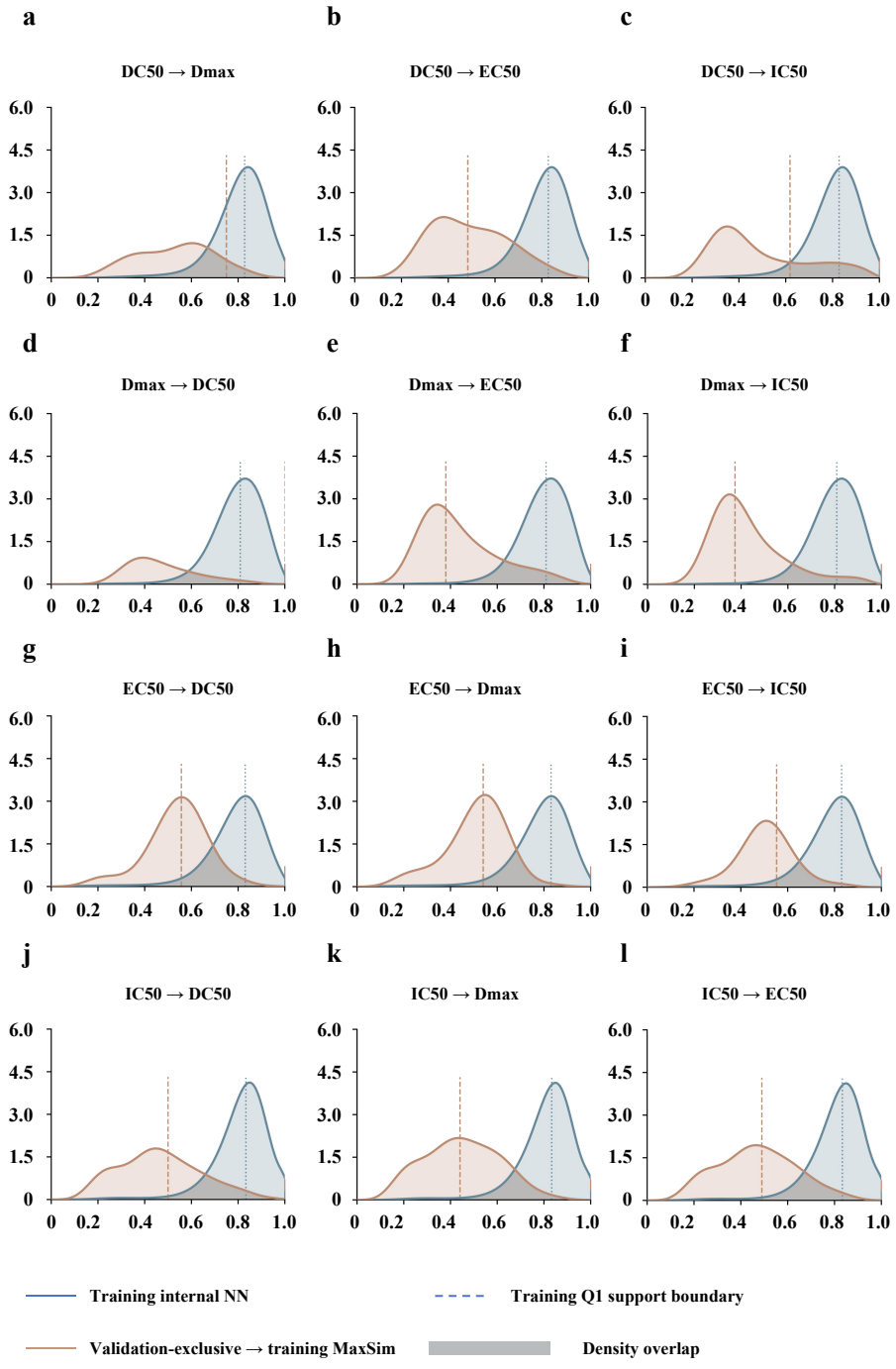


Figure B2: Cross-endpoint chemical structural support. Each panel compares the internal nearest-neighbor similarity distribution of the training endpoint (blue) with the maximum similarity of molecules exclusive to the validation endpoint to molecules in the training endpoint (orange). Panel titles indicate training → validation endpoint pairs. Dashed vertical lines mark the training distribution's first quartile (Q1 support boundary), and gray shading indicates density overlap.

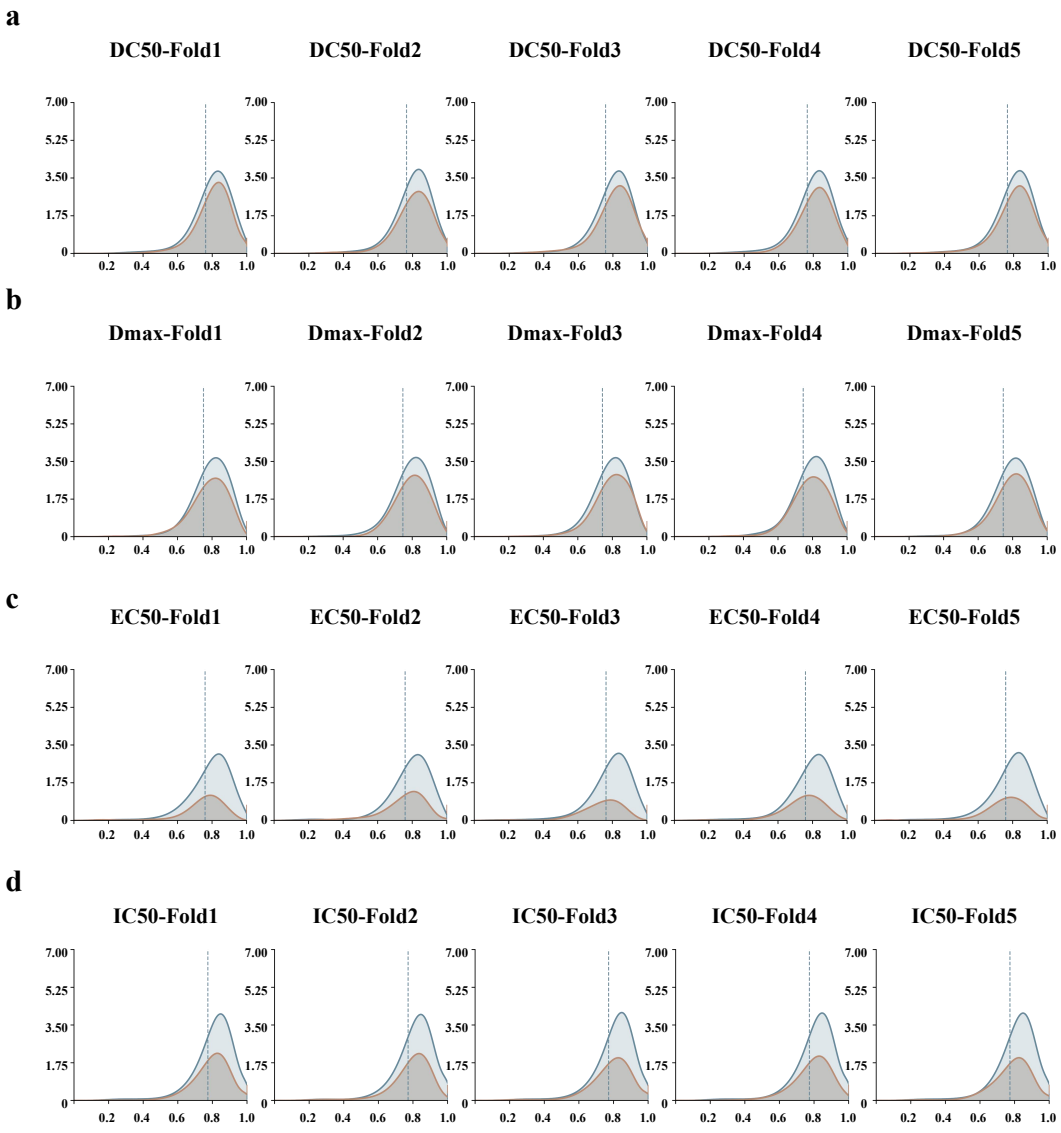


Figure B3: Molecular similarity under random five-fold cross-validation. Blue curves show within-training nearest-neighbor similarity; orange curves show maximum similarity to training molecules for molecules exclusive to validation. Similarity is measured using Morgan-fingerprint Tanimoto similarity. Dashed lines indicate the first quartile of the within-training nearest-neighbor similarity distribution. Rows (a–d) correspond to DC50, Dmax, EC50, and IC50; columns indicate folds.

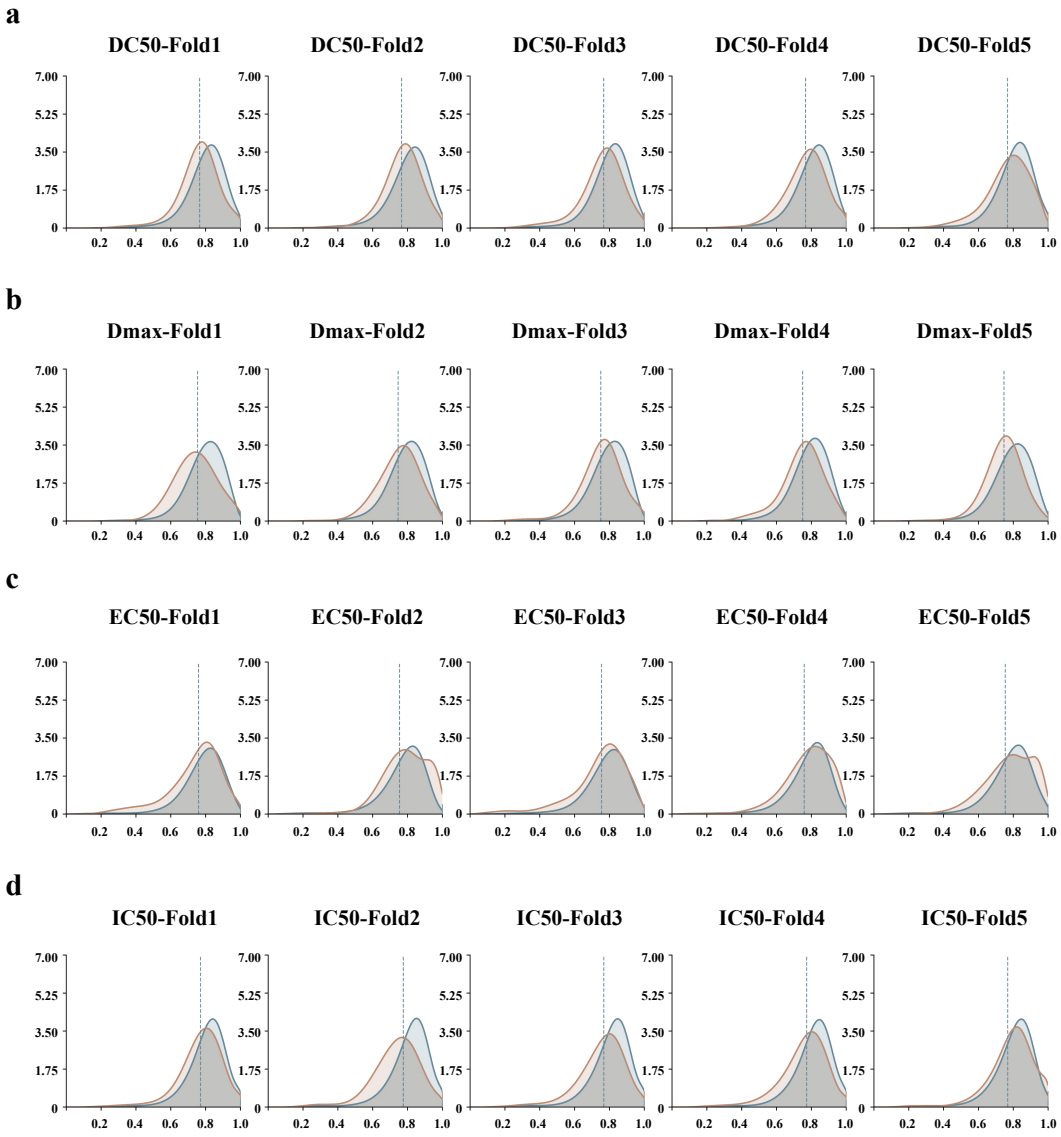


Figure B4: Molecular similarity under Murcko scaffold five-fold cross-validation. Blue curves show within-training nearest-neighbor similarity; orange curves show maximum validation-to-training similarity. Similarity is measured using Morgan-fingerprint Tanimoto similarity. Dashed lines indicate the first quartile of the within-training nearest-neighbor similarity distribution. Rows (a–d) correspond to DC50, Dmax, EC50, and IC50; columns indicate folds.

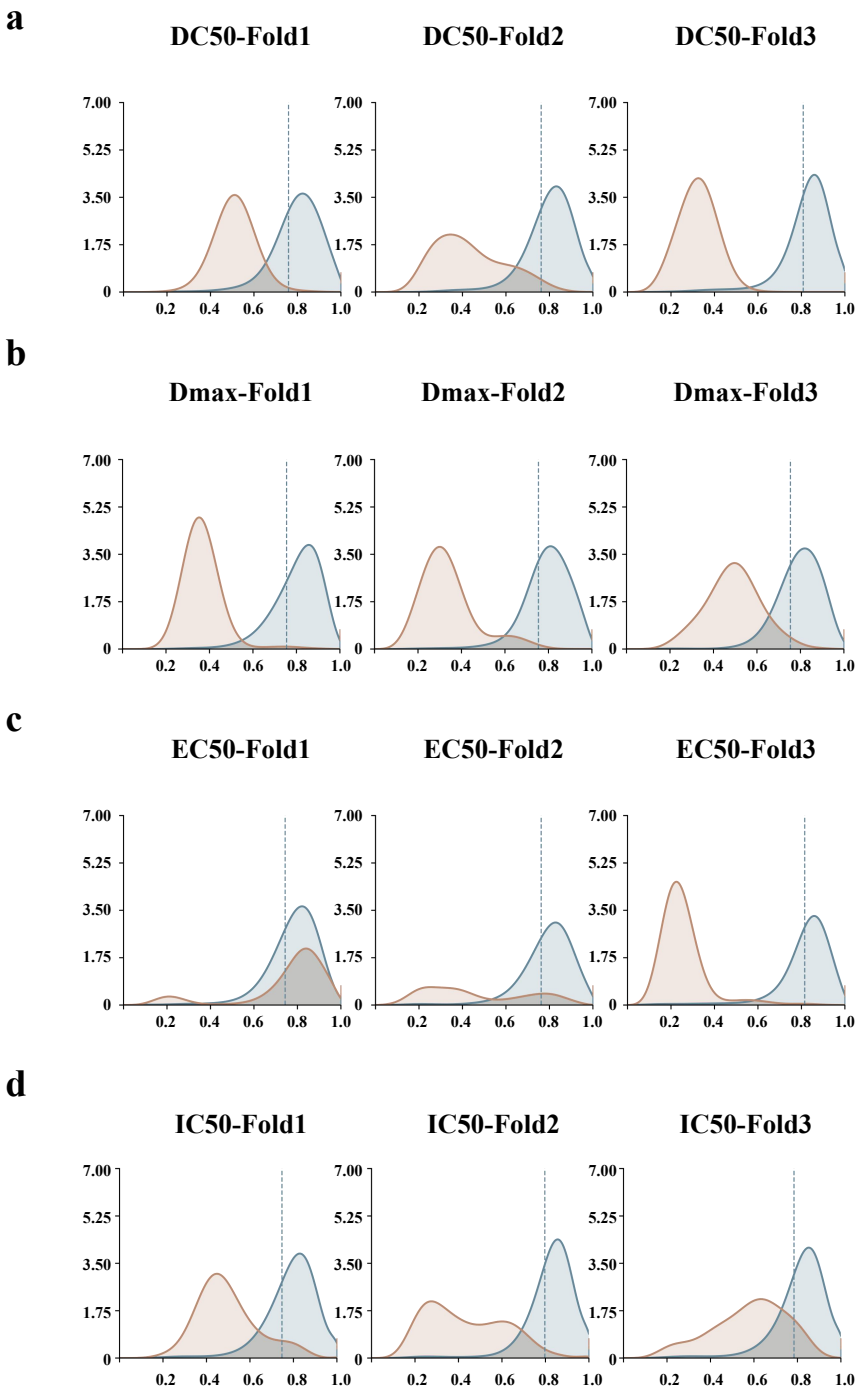


Figure B5: Molecular similarity under target-cluster three-fold cross-validation. Blue curves show within-training nearest-neighbor similarity; orange curves show maximum similarity to training molecules for molecules exclusive to validation. Similarity is measured using Morgan-fingerprint Tanimoto similarity. Dashed lines indicate the first quartile of the within-training nearest-neighbor similarity distribution. Rows (a–d) correspond to DC50, Dmax, EC50, and IC50; columns indicate folds.

Adaptor pretraining data. The pretraining pool comprised 12,380 unique molecule–target–effector triplets, obtained from 19,815 records spanning 16,898 quadruples; 9,574 triplets belonged to the unlabeled subset. Triplets were weighted equally and activity labels were hidden. For each validation fold, eligible pretraining inputs exclude held-out triplets across cell lines, together with held-out scaffolds or target clusters under the corresponding cold scheme. Target exclusion uses a cluster map spanning the full candidate pool. Negative examples are generated by replacing the target, effector, or both with entities drawn from the eligible pool, rejecting matches to observed eligible or held-out triplets. These pool counts precede fold-specific filtering. Sample assignments and processing code accompany the dataset.

C Model configurations and training details

C.1 Molecular and protein representations

We combined three molecular and three protein representations into nine pairs (Table C1). Target and effector proteins were encoded separately using the same representation type. All pretrained encoders were frozen.

Table C1: **Molecular and protein representations.** Dimensions refer to one molecule or one protein.

Representation	Dimension	Feature construction
Morgan	1,024	Binary Morgan fingerprint with radius 2.
MolDeBERTa	768	Final-layer CLS embedding of the SMILES input.
MuMo	768	First-token embedding after SMILES, 2D graph, BRICS, and 3D fusion.
AAC+DPC	420	Frequencies of 20 amino acids and 400 adjacent dipeptides.
ESM-2	1,280	Mean residue embedding from the 650M-parameter model.
ProstT5	1,024	Mean residue embedding in the amino-acid-to-fold direction.

Molecular features. Morgan fingerprints were computed from canonical isomeric SMILES using RDKit (Landrum et al., 2026) with chirality encoding disabled. MolDeBERTa used the `SaeedLab/MolDeBERTa-base-123M-mtr` checkpoint and the final hidden state of the CLS token, with a maximum input length of 128 tokens. MuMo used `zihaojing/MuMo-Pretrained` with the SMILES, molecular graph, BRICS, and geometry inputs produced by its graph preprocessing pipeline. The final first-token embedding represented each molecule, with a maximum SMILES length of 512 tokens. Molecules containing wildcard atoms were assigned zero coordinates for the geometry input.

Protein features. AAC frequencies were normalized by sequence length, and DPC frequencies by the number of adjacent residue pairs. ESM-2 used `facebook/esm2_t33_650M_UR50D`; ProstT5 used `Rostlab/ProstT5` in the amino-acid-to-fold direction. Residue embeddings were averaged to obtain one vector per protein. Sequences longer than 1,022 residues were divided into non-overlapping chunks, and chunk representations were combined by a residue-count-weighted mean.

Cell-line encoding and feature fusion. Cell lines were one-hot encoded using a vocabulary constructed from each training fold. The encoding included two distinct reserved categories: one for missing annotations and one shared by all cell lines unseen in the training fold. The baseline input was $x = [z_{\text{effector}} \| z_{\text{MGD}} \| z_{\text{target}} \| c_{\text{cell}}]$, where $\|$ denotes concatenation. Its dimension was $d_{\text{MGD}} + 2d_{\text{protein}} + d_{\text{cell}}$, where the three terms denote the molecular representation dimension, twice the protein representation dimension, and the cell vocabulary size, respectively.

C.2 Supervised models and training protocol

Each representation pair was evaluated with five model families, giving 45 configurations per task. Model settings (Table C2) were fixed across endpoints and splits. Classification and regression predictors were fitted separately on the Appendix B partitions; the joint DC50–Dmax task used the classification settings.

Table C2: **Supervised model configurations.** Settings apply to both classification and regression unless specified.

Model	Configuration
Random forest	300 trees; bootstrap sampling; $\sqrt{d}$ candidate features per split; unrestricted tree depth; minimum split size 2 and leaf size 1. Gini impurity for classification and squared error for regression.
RBF SVC/SVR	$C = 1$; $\gamma = 1/[d\,\mathrm{Var}(\mathbf{X}_{\mathrm{train}})]$; SVR $\epsilon = 0.1$.
MLP	Hidden layers of 256 and 64 units; ReLU activation and dropout 0.2 after each hidden layer; one output unit. AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 128, 50 epochs, and gradient norm clipping at 5.
XGBoost	250 boosting rounds; maximum depth 4; learning rate 0.05; minimum child weight 1; row and column sampling fractions 1; L_1 penalty 0 and L_2 penalty 1; histogram-based tree construction. Logistic objective for classification and squared error for regression.
TabPFN v3	Default v3 classification and regression checkpoints; estimator count initialized to 8 with automatic estimator scaling enabled; checkpoint-specific preprocessing and feature subsampling.

Preprocessing and optimization. SVC/SVR and MLP inputs were standardized using feature means and standard deviations fitted on the training fold. For SVR and MLP regression, target values were also standardized using training-fold statistics, and predictions were transformed back to the endpoint scale. Random forest and XGBoost operated on the concatenated features directly, while TabPFN used its built-in preprocessing. In Table C2, d denotes the predictor input dimension, and $\mathrm{Var}(\mathbf{X}_{\mathrm{train}})$ is the variance over all entries of the standardized training matrix.

The MLP used binary cross-entropy on logits for classification and mean squared error for regression, with shuffled mini-batches and final-epoch predictions. Supervised samples were equally weighted. Probability-based classifiers assigned high activity at a threshold of 0.5. SVC used decision scores for AUROC and AP and its native boundary for class predictions.

C.3 Association-based adaptor pretraining

For each endpoint, split, fold, and representation pair, we pretrained an adaptor on the Appendix B triplets with equal weight per triplet. Concatenated molecular and protein features were standardized using the pretraining pool. For input $u = [z_{\mathrm{effector}} \| z_{\mathrm{MGD}} \| z_{\mathrm{target}}]$, the adaptor mapped its standardized form $\widetilde{u}$ to a 512-dimensional vector,

$$z_{\mathrm{pre}} = g_\theta(\widetilde{u}) = \mathbf{W}_2\,\mathrm{ReLU}(\mathbf{W}_1\widetilde{u} + \mathbf{b}_1) + \mathbf{b}_2, \tag{C1}$$

with a hidden width of 1,024. A linear head produced a scalar triplet score $s(q) = \mathbf{w}^{\mathsf{T}} g_\theta(\widetilde{u}_q) + b$.

Triplet sampling and objective. For each observed triplet $q = (M, T, E)$, we sampled replacements of the target, effector, or both, keeping the molecule fixed. Replacement proteins were drawn according to their frequencies in the eligible pretraining pool. Candidates matching observed eligible or held-out triplets were rejected. One valid candidate of each replacement type was retained when available, giving a set $\mathcal{N}(q)$ of up to three sampled triplets. Candidates were resampled at each epoch. The loss was

$$\mathcal{L}_{\mathrm{pre}} = -\frac{1}{|\mathcal{B}|}\sum_{q\in\mathcal{B}} \log \frac{\exp(s(q)/\tau)}{\exp(s(q)/\tau) + \sum_{q^-\in\mathcal{N}(q)} \exp(s(q^-)/\tau)}, \tag{C2}$$

where $\mathcal{B}$ contains the observed triplets with at least one valid replacement in the mini-batch and $\tau = 0.1$.

Optimization and downstream prediction. The adaptor and scoring head were trained jointly for 60 epochs using AdamW, with learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 64, and gradient norm clipping at 5. After pretraining, the adaptor was frozen, and its output was concatenated with the cell-line encoding to form $x_{\mathrm{pre}} = [z_{\mathrm{pre}} \,\|\, c_{\mathrm{cell}}]$. The five downstream predictors were then fitted

using the configurations in Table C2. Classification and regression for a given endpoint shared the pretrained adaptor for the corresponding split, fold, and representation pair.

C.4 Supplementary analysis protocols

Supplementary analyses used the benchmark partitions and the model settings in Section C.2. Unless otherwise specified, supplementary analyses were conducted across all 45 baseline configurations. Association-pretraining comparisons used the adaptor described in Section C.3.

Training-set support. Chemical support was the maximum Morgan–Tanimoto similarity to a training molecule. We summarized support over unique held-out molecules absent from training and over all held-out samples, using 0.8 as the high-support threshold. Target support was the maximum value of $n_{\mathrm{ident}}/\max(L_q, L_t)$ from MMseqs2 alignments to training targets, where n_{ident} counts identical aligned residues and L_q, L_t are full sequence lengths. Chemical support was binned at 0.4, 0.6, and 0.8. AUROC was decomposed by the bins of positive–negative sample pairs. Each category contributed its pair fraction multiplied by its conditional AUROC; its contribution above chance used conditional AUROC minus 0.5.

Nearest-neighbor prediction. Nearest training molecules were retrieved by Morgan–Tanimoto similarity or by cosine similarity for learned embeddings. Training observations were averaged per molecule: classification scores were the mean binary label, and regression values were the mean endpoint-scale target. The nearest molecule's value was assigned to the query, with similarity ties resolved by sorted molecule identifier.

Protein-feature controls. Protein embeddings were replaced by fixed Gaussian vectors of matching dimension or permuted across protein identities, with each identity assigned a different donor embedding. Gaussian coordinates were scaled to the means and standard deviations of unique proteins in the supervised training set and eligible association pool. Identical sequences shared the same replacement in target and effector roles.

Evidence quality and cellular information. Type 1 samples were reconstructed by averaging Type 1 measurements within each molecule–target–effector–cell combination and retaining the benchmark fold assignments. Type 1-trained and Type 1+2-trained models were evaluated on the same Type 1 samples using Type 1-derived labels and values. For cell-feature ablation, models were refitted after removing the cell one-hot vector, retaining the same samples, labels, and folds. Cell-annotation analyses grouped held-out predictions by reported versus missing cell identity and calculated AUROC and class-specific recall within each group.

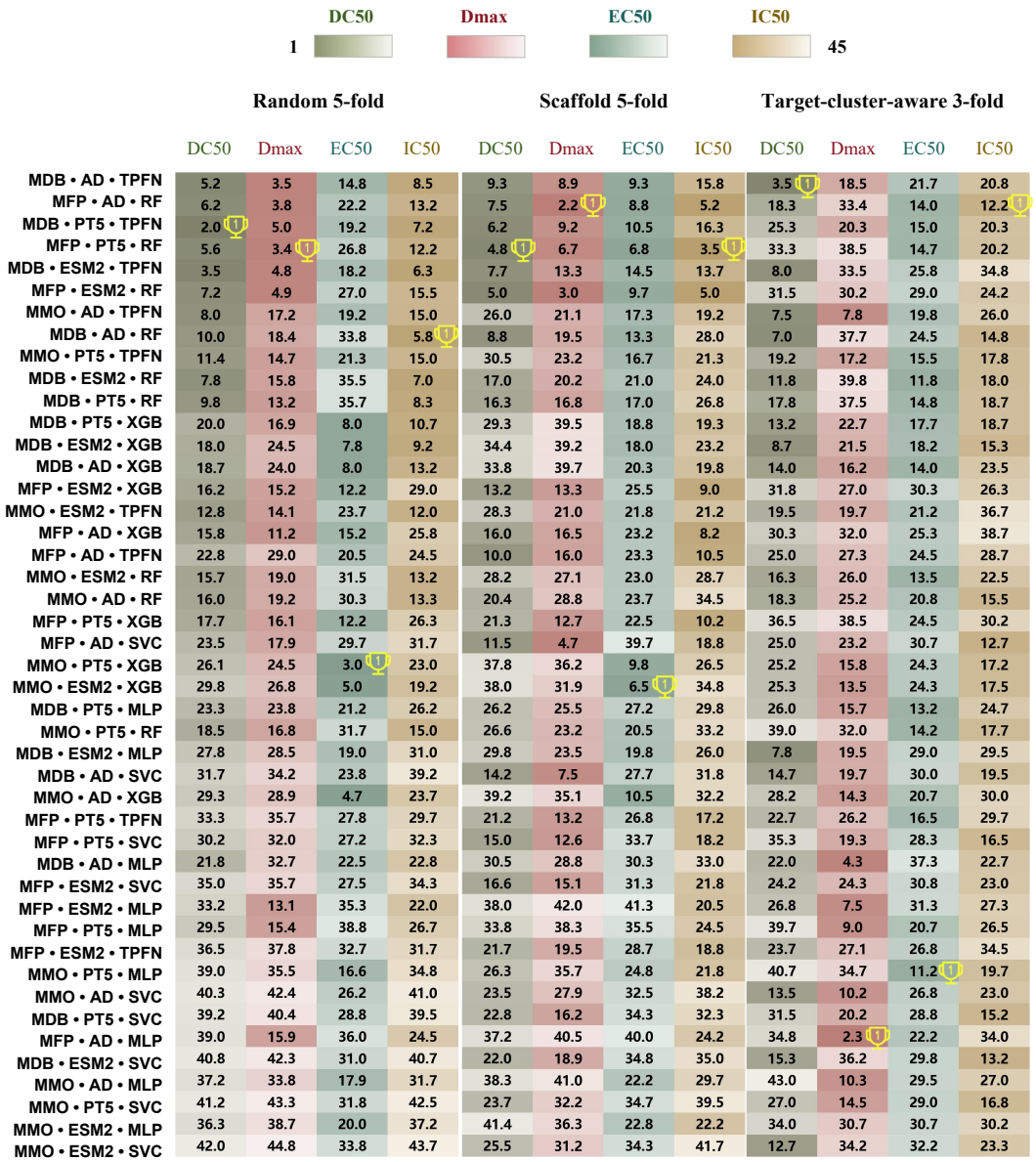
Within-context molecular ranking. Within each held-out fold, molecules were grouped by target, effector, and reported cell line. Groups contained at least five molecules and two distinct measured activity values. Pairwise ordering accuracy excluded equal-activity pairs and assigned 0.5 to prediction ties. Scores were averaged across eligible folds for each background, then equally across backgrounds and configurations.

Joint degradation criteria. The joint classifier's scores were evaluated by comparing compounds meeting both DC50 $\leq$ 100 nM and Dmax $\geq$ 80% against compounds meeting only the DC50 criterion or only the Dmax criterion. AUROC was calculated separately for the two comparisons within each fold and configuration.

Confidence filtering and calibration. Each outer training set was randomly divided into 80% fitting and 20% calibration samples (partition seed 2026); predictors used seed 42. SVC scores were mapped to probabilities by Platt scaling on the calibration set. Predictions were ranked by binary entropy $H(p) = -p\log p - (1-p)\log(1-p)$, and error rates were compared before and after removing the highest-entropy 20%. Temperature scaling minimized calibration-set negative log-likelihood over $\log T \in [-5, 5]$. High-confidence errors were incorrect predictions with $\max(p, 1-p) \geq 0.9$.

C.5 IMPLEMENTATION

Supervised fitting and adaptor pretraining used random seed 42. The training environment comprised Python 3.12.3, scikit-learn 1.6.1, XGBoost 2.1.3, TabPFN 8.1.0, and PyTorch 2.7.0a0 (NVIDIA 25.03 build), with NVIDIA GeForce RTX 4090 GPUs. Molecular and protein embedding extraction used PyTorch 2.7.1; Transformers versions were 4.49.0 for MolDeBERTa and 4.46.3 for ESM-2, ProstT5, and MuMo. Morgan fingerprints were generated with RDKit 2025.03.5.



	Random 5-fold				Scaffold 5-fold				Target-cluster-aware 3-fold			
	DC50	Dmax	EC50	IC50	DC50	Dmax	EC50	IC50	DC50	Dmax	EC50	IC50
MDB • AD • TPFN	5.2	3.5	14.8	8.5	9.3	8.9	9.3	15.8	3.5	18.5	21.7	20.8
MFP • AD • RF	6.2	3.8	22.2	13.2	7.5	2.2	8.8	5.2	18.3	33.4	14.0	12.2
MDB • PT5 • TPFN	2.0	5.0	19.2	7.2	6.2	9.2	10.5	16.3	25.3	20.3	15.0	20.3
MFP • PT5 • RF	5.6	3.4	26.8	12.2	4.8	6.7	6.8	3.5	33.3	38.5	14.7	20.2
MDB • ESM2 • TPFN	3.5	4.8	18.2	6.3	7.7	13.3	14.5	13.7	8.0	33.5	25.8	34.8
MFP • ESM2 • RF	7.2	4.9	27.0	15.5	5.0	3.0	9.7	5.0	31.5	30.2	29.0	24.2
MMO • AD • TPFN	8.0	17.2	19.2	15.0	26.0	21.1	17.3	19.2	7.5	7.8	19.8	26.0
MDB • AD • RF	10.0	18.4	33.8	5.8	8.8	19.5	13.3	28.0	7.0	37.7	24.5	14.8
MMO • PT5 • TPFN	11.4	14.7	21.3	15.0	30.5	23.2	16.7	21.3	19.2	17.2	15.5	17.8
MDB • ESM2 • RF	7.8	15.8	35.5	7.0	17.0	20.2	21.0	24.0	11.8	39.8	11.8	18.0
MDB • PT5 • RF	9.8	13.2	35.7	8.3	16.3	16.8	17.0	26.8	17.8	37.5	14.8	18.7
MDB • PT5 • XGB	20.0	16.9	8.0	10.7	29.3	39.5	18.8	19.3	13.2	22.7	17.7	18.7
MDB • ESM2 • XGB	18.0	24.5	7.8	9.2	34.4	39.2	18.0	23.2	8.7	21.5	18.2	15.3
MDB • AD • XGB	18.7	24.0	8.0	13.2	33.8	39.7	20.3	19.8	14.0	16.2	14.0	23.5
MFP • ESM2 • XGB	16.2	15.2	12.2	29.0	13.2	13.3	25.5	9.0	31.8	27.0	30.3	26.3
MMO • ESM2 • TPFN	12.8	14.1	23.7	12.0	28.3	21.0	21.8	21.2	19.5	19.7	21.2	36.7
MFP • AD • XGB	15.8	11.2	15.2	25.8	16.0	16.5	23.2	8.2	30.3	32.0	25.3	38.7
MFP • AD • TPFN	22.8	29.0	20.5	24.5	10.0	16.0	23.3	10.5	25.0	27.3	24.5	28.7
MMO • ESM2 • RF	15.7	19.0	31.5	13.2	28.2	27.1	23.0	28.7	16.3	26.0	13.5	22.5
MMO • AD • RF	16.0	19.2	30.3	13.3	20.4	28.8	23.7	34.5	18.3	25.2	20.8	15.5
MFP • PT5 • XGB	17.7	16.1	12.2	26.3	21.3	12.7	22.5	10.2	36.5	38.5	24.5	30.2
MFP • AD • SVC	23.5	17.9	29.7	31.7	11.5	4.7	39.7	18.8	25.0	23.2	30.7	12.7
MMO • PT5 • XGB	26.1	24.5	3.0	23.0	37.8	36.2	9.8	26.5	25.2	15.8	24.3	17.2
MMO • ESM2 • XGB	29.8	26.8	5.0	19.2	38.0	31.9	6.5	34.8	25.3	13.5	24.3	17.5
MDB • PT5 • MLP	23.3	23.8	21.2	26.2	26.2	25.5	27.2	29.8	26.0	15.7	13.2	24.7
MMO • PT5 • RF	18.5	16.8	31.7	15.0	26.6	23.2	20.5	33.2	39.0	32.0	14.2	17.7
MDB • ESM2 • MLP	27.8	28.5	19.0	31.0	29.8	23.5	19.8	26.0	7.8	19.5	29.0	29.5
MDB • AD • SVC	31.7	34.2	23.8	39.2	14.2	7.5	27.7	31.8	14.7	19.7	30.0	19.5
MMO • AD • XGB	29.3	28.9	4.7	23.7	39.2	35.1	10.5	32.2	28.2	14.3	20.7	30.0
MFP • PT5 • TPFN	33.3	35.7	27.8	29.7	21.2	13.2	26.8	17.2	22.7	26.2	16.5	29.7
MFP • PT5 • SVC	30.2	32.0	27.2	32.3	15.0	12.6	33.7	18.2	35.3	19.3	28.3	16.5
MDB • AD • MLP	21.8	32.7	22.5	22.8	30.5	28.8	30.3	33.0	22.0	4.3	37.3	22.7
MFP • ESM2 • SVC	35.0	35.7	27.5	34.3	16.6	15.1	31.3	21.8	24.2	24.3	30.8	23.0
MFP • ESM2 • MLP	33.2	13.1	35.3	22.0	38.0	42.0	41.3	20.5	26.8	7.5	31.3	27.3
MFP • PT5 • MLP	29.5	15.4	38.8	26.7	33.8	38.3	35.5	24.5	39.7	9.0	20.7	26.5
MFP • ESM2 • TPFN	36.5	37.8	32.7	31.7	21.7	19.5	28.7	18.8	23.7	27.1	26.8	34.5
MMO • PT5 • MLP	39.0	35.5	16.6	34.8	26.3	35.7	24.8	21.8	40.7	34.7	11.2	19.7
MMO • AD • SVC	40.3	42.4	26.2	41.0	23.5	27.9	32.5	38.2	13.5	10.2	26.8	23.0
MDB • PT5 • SVC	39.2	40.4	28.8	39.5	22.8	16.2	34.3	32.3	31.5	20.2	28.8	15.2
MFP • AD • MLP	39.0	15.9	36.0	24.5	37.2	40.5	40.0	24.2	34.8	2.3	22.2	34.0
MDB • ESM2 • SVC	40.8	42.3	31.0	40.7	22.0	18.9	34.8	35.0	15.3	36.2	29.8	13.2
MMO • AD • MLP	37.2	33.8	17.9	31.7	38.3	41.0	22.2	29.7	43.0	10.3	29.5	27.0
MMO • PT5 • SVC	41.2	43.3	31.8	42.5	23.7	32.2	34.7	39.5	27.0	14.5	29.0	16.8
MMO • ESM2 • MLP	36.3	38.7	20.0	37.2	41.4	36.3	22.8	22.2	34.0	30.7	30.7	30.2
MMO • ESM2 • SVC	42.0	44.8	33.8	43.7	25.5	31.2	34.3	41.7	12.7	34.2	32.2	23.3

Figure D1: Complete configuration rankings across endpoints and evaluation splits. Each cell reports the mean rank of one of the 45 molecular representation–protein representation–classifier configurations, calculated from AUROC, AP, F1, recall, accuracy, and precision within each endpoint and split. Lower values indicate better aggregate classification performance, with ranks ranging from 1 to 45. MDB: MolDeBERTa; MFP: Morgan fingerprint; MMO: MuMo; AD: AAC+DPC; ESM2: ESM-2; PT5: ProstT5; TPFN: TabPFN v3; RF: random forest; XGB: XGBoost; SVC: RBF support vector classifier; MLP: multilayer perceptron.

D Additional Experimental Results

For each endpoint and split, we ranked 45 configurations separately by AUROC, average precision (AP), F1, recall, accuracy, and precision, assigning average ranks to ties. The arithmetic mean of these six ranks determined the top five configurations. Ties in mean rank were resolved by higher AUROC. Tables D1–D3 report their cross-validation fold means. Each configuration combines a molecular representation, a protein representation, and a predictive model.

Table D1: Top five configurations per endpoint under random five-fold cross-validation. Mean rank averages the ranks of AUROC, AP, F1, recall, accuracy, and precision across 45 configurations; lower values indicate better performance. Metrics are cross-validation fold means. Bold entries mark the highest unrounded value among the five displayed configurations for each endpoint.

Configuration	Mean Rank ↓	AUROC	AP	F1	Recall	Acc.	Prec.
DC50							
MolDeBERTa / ProstT5 / TabPFN	2.00	**0.886**	0.903	**0.832**	**0.854**	**0.811**	0.811
MolDeBERTa / ESM2 / TabPFN	3.50	0.884	0.902	0.829	0.851	0.808	0.807
MolDeBERTa / AAC / TabPFN	5.17	0.886	**0.903**	0.823	0.840	0.803	0.808
Morgan / ProstT5 / RF	5.58	0.875	0.885	0.825	0.837	0.805	0.813
Morgan / AAC / RF	6.25	0.876	0.883	0.824	0.835	0.805	**0.814**
Dmax							
Morgan / ProstT5 / RF	3.42	0.895	0.893	**0.827**	**0.855**	**0.817**	0.801
MolDeBERTa / AAC / TabPFN	3.50	**0.899**	**0.899**	0.822	0.840	0.814	**0.806**
Morgan / AAC / RF	3.83	0.895	0.894	0.826	0.853	0.816	0.801
MolDeBERTa / ESM2 / TabPFN	4.83	0.897	0.899	0.821	0.843	0.811	0.800
Morgan / ESM2 / RF	4.92	0.897	0.894	0.824	**0.855**	0.813	0.795
EC50							
MuMo / ProstT5 / XGB	3.00	**0.954**	**0.955**	**0.904**	**0.913**	**0.893**	**0.895**
MuMo / AAC / XGB	4.67	0.952	0.954	0.902	0.913	0.892	0.893
MuMo / ESM2 / XGB	5.00	0.954	0.954	0.901	0.911	0.890	0.892
MolDeBERTa / ESM2 / XGB	7.83	0.950	0.955	0.890	0.894	0.879	0.887
MolDeBERTa / AAC / XGB	8.00	0.950	0.955	0.890	0.893	0.879	0.887
IC50							
MolDeBERTa / AAC / RF	5.83	0.931	0.858	**0.796**	**0.784**	**0.885**	0.809
MolDeBERTa / ESM2 / TabPFN	6.33	0.935	0.859	0.783	0.756	0.880	0.814
MolDeBERTa / ESM2 / RF	7.00	0.930	0.858	0.791	0.777	0.882	0.807
MolDeBERTa / ProstT5 / TabPFN	7.17	**0.936**	**0.861**	0.782	0.752	0.880	**0.815**
MolDeBERTa / ProstT5 / RF	8.33	0.929	0.858	0.788	0.771	0.881	0.808

Configuration: molecular / protein / predictive model. Morgan: Morgan-1024; AAC: AAC+DPC; ESM2: ESM-2-1280; RF: random forest; XGB: XGBoost; SVC: RBF SVC; TabPFN: TabPFN v3; MLP: multilayer perceptron.

Table D2: Top five configurations per endpoint under scaffold five-fold cross-validation. Mean rank averages the ranks of AUROC, AP, F1, recall, accuracy, and precision across 45 configurations; lower values indicate better performance. Metrics are cross-validation fold means. Bold entries mark the highest unrounded value among the five displayed configurations for each endpoint.

Configuration	Mean Rank ↓	AUROC	AP	F1	Recall	Acc.	Prec.
DC50							
Morgan / ProstT5 / RF	4.83	0.821	0.846	**0.786**	0.806	**0.759**	**0.770**
Morgan / ESM2 / RF	5.00	**0.823**	**0.848**	0.785	0.806	0.758	0.770
MolDeBERTa / ProstT5 / TabPFN	6.17	0.816	0.846	0.781	0.808	0.752	0.757
Morgan / AAC / RF	7.50	0.820	0.844	0.781	0.800	0.754	0.769
MolDeBERTa / ESM2 / TabPFN	7.67	0.814	0.842	0.778	**0.812**	0.746	0.748
Dmax							
Morgan / AAC / RF	2.17	0.822	0.833	**0.763**	0.783	**0.752**	**0.748**
Morgan / ESM2 / RF	3.00	**0.824**	**0.837**	0.759	0.782	0.747	0.743
Morgan / AAC / SVC	4.67	0.820	0.823	0.761	**0.785**	0.751	0.745
Morgan / ProstT5 / RF	6.67	0.819	0.829	0.755	0.773	0.744	0.743
MolDeBERTa / AAC / SVC	7.50	0.819	0.824	0.756	0.780	0.745	0.739
EC50							
MuMo / ESM2 / XGB	6.50	0.929	0.932	**0.876**	0.879	**0.863**	**0.876**
Morgan / ProstT5 / RF	6.83	0.936	**0.939**	0.874	0.889	0.859	0.864
Morgan / AAC / RF	8.83	**0.937**	0.939	0.871	0.887	0.856	0.862
MolDeBERTa / AAC / TabPFN	9.33	0.928	0.931	0.874	0.888	0.860	0.866
Morgan / ESM2 / RF	9.67	0.936	0.938	0.871	**0.890**	0.856	0.858
IC50							
Morgan / ProstT5 / RF	3.50	**0.865**	0.721	**0.635**	**0.569**	**0.814**	0.727
Morgan / ESM2 / RF	5.00	0.864	**0.721**	0.627	0.560	0.812	0.723
Morgan / AAC / RF	5.17	0.862	0.718	0.630	0.564	0.812	0.724
Morgan / AAC / XGB	8.17	0.854	0.719	0.608	0.517	0.810	**0.747**
Morgan / ESM2 / XGB	9.00	0.854	0.715	0.610	0.525	0.809	0.740

Configuration: molecular / protein / predictive model. Morgan: Morgan-1024; AAC: AAC+DPC; ESM2: ESM-2-1280; RF: random forest; XGB: XGBoost; SVC: RBF SVC; TabPFN: TabPFN v3; MLP: multilayer perceptron.

Table D3: Top five configurations per endpoint under target-cluster three-fold cross-validation. Mean rank averages the ranks of AUROC, AP, F1, recall, accuracy, and precision across 45 configurations; lower values indicate better performance. Metrics are cross-validation fold means. Bold entries mark the highest unrounded value among the five displayed configurations for each endpoint.

Configuration	Mean Rank ↓	AUROC	AP	F1	Recall	Acc.	Prec.
DC50							
MolDeBERTa / AAC / TabPFN	3.50	**0.700**	**0.722**	**0.714**	0.833	**0.648**	0.626
MolDeBERTa / AAC / RF	7.00	0.643	0.668	0.695	0.760	0.648	**0.646**
MuMo / AAC / TabPFN	7.50	0.644	0.668	0.696	0.864	0.597	0.584
MolDeBERTa / ESM2 / MLP	7.83	0.696	0.716	0.673	0.754	0.621	0.629
MolDeBERTa / ESM2 / TabPFN	8.00	0.646	0.687	0.703	**0.935**	0.575	0.564
Dmax							
Morgan / AAC / MLP	2.33	**0.589**	**0.632**	**0.497**	0.380	**0.631**	**0.729**
MolDeBERTa / AAC / MLP	4.33	0.579	0.607	0.480	**0.403**	0.610	0.596
Morgan / ESM2 / MLP	7.50	0.558	0.601	0.426	0.361	0.569	0.615
MuMo / AAC / TabPFN	7.83	0.587	0.588	0.368	0.273	0.571	0.666
Morgan / ProstT5 / MLP	9.00	0.588	0.606	0.318	0.248	0.557	0.652
EC50							
MuMo / ProstT5 / MLP	11.17	**0.762**	**0.693**	0.544	0.552	0.789	0.713
MolDeBERTa / ESM2 / RF	11.83	0.628	0.628	0.554	0.490	**0.820**	**0.856**
MolDeBERTa / ProstT5 / MLP	13.17	0.716	0.624	0.553	**0.561**	0.790	0.695
MuMo / ESM2 / RF	13.50	0.634	0.621	0.550	0.487	0.815	0.851
MolDeBERTa / AAC / XGB	14.00	0.641	0.620	**0.556**	0.536	0.795	0.786
IC50							
Morgan / AAC / RF	12.17	0.499	0.346	**0.316**	0.381	0.713	0.575
Morgan / AAC / SVC	12.67	0.594	0.382	0.227	0.152	**0.775**	**0.766**
MolDeBERTa / ESM2 / SVC	13.17	0.515	0.323	0.279	0.374	0.716	0.614
MolDeBERTa / AAC / RF	14.83	0.543	**0.390**	0.197	0.172	0.749	0.574
MolDeBERTa / ProstT5 / SVC	15.17	**0.606**	0.333	0.253	**0.405**	0.619	0.498

Configuration: molecular / protein / predictive model. Morgan: Morgan-1024; AAC: AAC+DPC; ESM2: ESM-2-1280; RF: random forest; XGB: XGBoost; SVC: RBF SVC; TabPFN: TabPFN v3; MLP: multilayer perceptron.

Table D4: Selected generalization controls and association-pretraining results.

A. Nearest-neighbor prediction

Split	Supervised AUROC	1-NN AUROC
Random	0.895	0.804
Target-cluster	0.561	0.571

B. Association-based adaptor: Joint task, target-cluster split

Metric	Raw	With adaptor	Δ
AUROC	0.481	0.522	+0.041
AP	0.499	0.531	+0.032

Table D5: Effects of evidence precision and cell-line information across evaluation splits.

Comparison	Random	Scaffold	Target-cluster
Type 1 vs. Type 1+2, ΔAUROC	+0.007	+0.007	+0.024
Type 1 vs. Type 1+2, ΔR^2	+0.026	+0.037	+0.183
With vs. without cell features, ΔR^2	+0.018	+0.020	+0.048
With vs. without cell features (IC50), ΔR^2	+0.024	+0.047	+0.118

Table D6: Selected results for molecular ranking, joint degradation criteria, and confidence-based filtering.

A. Within-context molecular ranking: target-cluster split

Endpoint	Pairwise ordering accuracy
DC50	0.532
Dmax	0.499
EC50	0.647
IC50	0.534

B. Joint degradation criteria: scaffold split

Comparison	AUROC
Both vs. DC50-only	0.840
Both vs. Dmax-only	0.630

C. Confidence filtering and calibration

Analysis	Random	Scaffold	Target-cluster
Error reduction after 20% entropy filtering (pp)	5.3–6.0	4.5–5.1	0.8–1.7
High-confidence errors, MLP (%)	–	–	24.9 → 9.0

Table D7: Component-level performance across evaluation splits.

		Random			Scaffold			Target-cluster		
Type	Component	MR↓	AUROC↑	R^2↑	MR↓	AUROC↑	R^2↑	MR↓	AUROC↑	R^2↑
MGD repr.	Morgan-1024	23.49	.895	**.656**	**19.02**	**.844**	**.509**	25.98	.549	−.213
	MolDeBERTa	**20.18**	**.898**	.650	22.12	.839	.500	**20.52**	**.575**	**−.177**
	MuMo	25.33	.891	.631	27.86	.834	.487	22.50	.558	−.179
Protein repr.	AAC+DPC	**21.94**	**.896**	**.652**	**22.59**	.839	**.500**	**21.64**	.556	**−.144**
	ESM-2	24.01	.893	.642	23.68	.839	.497	24.15	.547	−.189
	ProstT5	23.05	.895	.643	22.72	**.840**	**.500**	23.21	**.578**	−.236
Model	RF	**16.37**	**.907**	.655	**17.08**	.848	.503	22.63	.559	**−.088**
	MLP	28.02	.888	.649	30.80	.823	.480	24.31	.541	−.317
	SVC/SVR	35.05	.872	.596	25.60	.835	.490	23.07	**.593**	−.127
	TabPFN v3	17.90	.905	**.677**	17.48	**.850**	**.528**	**21.89**	.570	−.094
	XGBoost	17.66	.901	.651	24.04	.839	.494	23.09	.540	−.323

Note: Component means are equally averaged across four endpoints; MR denotes the six-metric classification marginal rank.