CHEMICAL-SUPPORT-RESOLVED EVALUATION OF MODEL-SELECTION STABILITY UNDER MOLECULAR EXTRAPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular benchmarks are commonly used not only to estimate predictive performance, but also to select which model should be deployed. This practice implicitly assumes that the model ordering observed on a benchmark remains informative when target compounds move beyond the chemical support of the training data. We examine this assumption through chemical-support-resolved evaluation, which characterizes each held-out drug by its maximum similarity to the training collection and evaluates both predictive performance and model-order stability as support decreases. Across RNA–drug, drug–protein, and drug–microbe association tasks, together with a separate ChEMBL37 direct-human drug–protein evaluation, we find that conventional cold-start partitions do not define a uniform extrapolation regime: even under Scaffold-Cold, 8.2–15.3% of test drugs retain relatively close training analogues. Predictive degradation and model-order stability are distinct. In the extreme-support regime ($S_{\max} < 0.2$), the controlled model set exhibits substantial ranking transitions in Yamanishi_08, MDAD, and ChEMBL37; in Yamanishi_08, Random Pair and low-support rankings yield Spearman $\rho = -0.302$ and 48.8% near-tie-adjusted pairwise reversals. Equal-size moderate-support controls show significantly greater ranking agreement, indicating that the transition is not explained by sparse cohort size alone. These results show that a benchmark leaderboard should be interpreted relative to the chemical-support regime in which model selection will be applied, rather than assumed to remain informative across molecular extrapolation.

1 INTRODUCTION

Benchmarks are used not only to measure predictive performance, but also to compare candidate models and inform model selection for real-world deployment. Under out-of-distribution (OOD) conditions, however, the model ranked first on familiar data may not remain first in a less supported regime, because different models can fail at different rates. This means that distribution-shift evaluation should ask not only whether performance decreases, but also whether the model ranking itself remains stable. The latter is essential if benchmark results are to guide model selection (Koh et al., 2021; Miller et al., 2021).

This distinction is particularly important in molecular machine learning, where test compounds can differ substantially in how well they are represented within the chemical space covered by the training data. Standard evaluation protocols typically address this issue through Random Pair, Drug-Cold, or Scaffold-Cold splits. Similarity Aware Evaluation (SAE) also relates evaluation to molecular similarity (Zhang et al., 2025). These protocols are useful because they control different forms of entity- or scaffold-level overlap, yet the resulting split labels remain only coarse indicators of chemical novelty. Compounds assigned to the same split can differ markedly in their similarity to the closest training analogues, while compounds from different split types may still occupy nearby regions of chemical space. Consequently, split membership alone does not provide a calibrated measure of how well a test compound is supported by the training chemistry.

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.

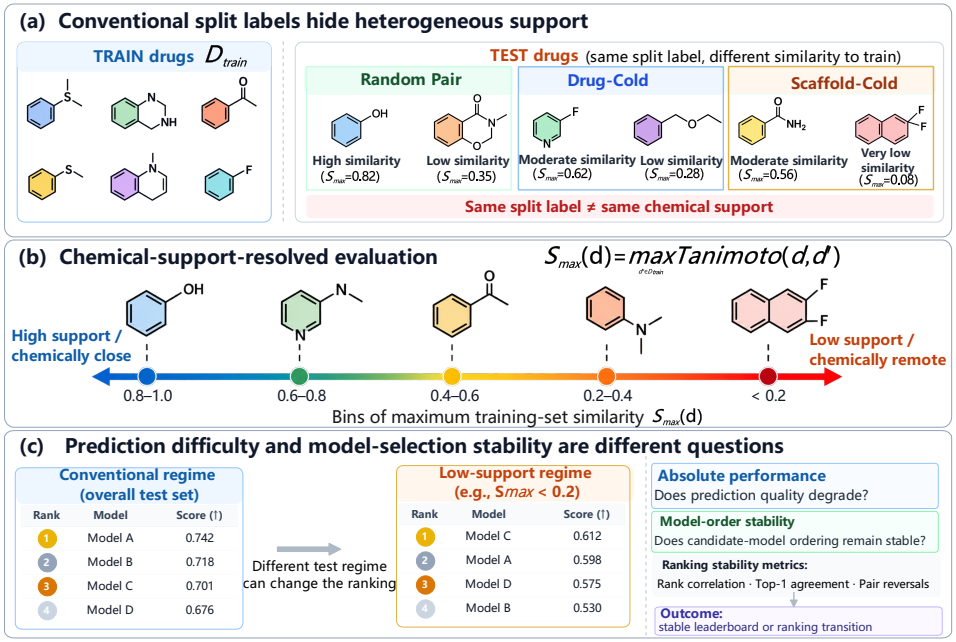


Figure 1: **Conceptual overview of chemical-support-resolved evaluation.** (a) Conventional split labels can contain chemically heterogeneous test drugs. (b) $S_{\max}$ provides a continuous measure of training-set chemical support. (c) The framework distinguishes absolute performance degradation from model-order stability and diagnoses whether a conventional leaderboard remains informative under low support.

We make this notion of support explicit by characterizing each held-out drug according to its maximum similarity to the training collection,

$$S_{\max}(d) = \max_{d' \in \mathcal{D}_{\text{train}}} \text{Tanimoto}(d, d'),$$

computed using a pre-specified molecular fingerprint. Lower $S_{\max}$ corresponds to weaker local support from the training chemical space. This definition provides a continuous description of the support heterogeneity left unresolved by coarse split labels. We ask whether the *relative ordering* of candidate models remains stable as chemical support changes.

This distinction matters because overall performance and model ranking can change in different ways. As chemical support weakens, all models may perform worse while still preserving the same ranking, in which case a conventional leaderboard remains useful for model selection. However, model-specific performance curves may also cross, even when the overall performance drop is small, meaning that the best-performing model can depend on the support regime. The key question is therefore not only whether extrapolation becomes more difficult, but also whether the evaluation protocol provides a stable basis for choosing among competing models.

Figure 1 summarizes the proposed evaluation framework. Organizing held-out compounds by the continuous support variable $S_{\max}$ exposes heterogeneity within conventional split labels and separates predictive performance from model-order stability across support regimes. We quantify the latter using rank agreement, top-model consistency, and near-tie-adjusted pairwise reversals.

We evaluate model-order stability across molecular-association settings with distinct label semantics, analyzing each dataset separately under a common drug-side definition of training-set chemical support. A controlled eight-model set provides the primary comparison, complemented by additional interaction-prediction architectures and a separate ChEMBL37 direct-human drug–target interaction (DTI) evaluation. Robustness is assessed through within-Drug-Cold comparisons, matched-support analyses, finite-cohort uncertainty estimates, alternative molecular representations, and label-disjoint model selection.

Table 1: SAE and this work target different evaluation objects.

	Similarity-aware evaluation	Our work
Object	Individual-model performance	Candidate-model ordering
Question	Does performance degrade?	Does candidate-model ordering remain stable?
Output	Similarity–performance curve	Rank agreement and meaningful pairwise reversals
Consequence	Support-specific performance	Oracle gap and label-disjoint selection simulation

1.1 Our Contributions

Chemical-support-resolved evaluation for model selection. We characterize each held-out compound by its continuous chemical support from the training set rather than relying solely on coarse split labels such as Drug-Cold or Scaffold-Cold. This formulation separates predictive degradation from model-order stability, revealing whether benchmark rankings remain informative as target compounds move into progressively less supported regions of chemical space.

A controlled evaluation of leaderboard stability under chemical extrapolation. We examine how model rankings change as chemical support decreases and whether these transitions persist when evaluation protocol, near ties, and finite-cohort effects are explicitly controlled. We combine rank-based stability measures with within-Drug-Cold comparisons, equal-size moderate-support controls, and whole-drug uncertainty analyses to distinguish support-conditioned changes in model ordering from variation associated with split differences and sparse low-support cohorts.

Cross-task evidence reveals conditional, rather than universal, ranking instability. Across multiple molecular-association settings and a separate ChEMBL37 evaluation, lower chemical support is associated with heterogeneous changes in model ordering rather than a uniform ranking response. Several controlled settings show substantial low-support ranking transitions, whereas other tasks and additional interaction-prediction architectures exhibit weaker or more stable ordering changes. These results indicate that ranking instability is context dependent, varying with the task, candidate model set, and molecular representation used to define chemical support.

2 Related Work

Molecular evaluation has increasingly moved beyond random pair splits toward protocols that control structural overlap and better expose distribution shift. Similarity to the training collection has long served as a local indicator of prediction reliability (Sheridan et al., 2004), with extended-connectivity fingerprints, Tanimoto similarity, and Bemis–Murcko scaffolds providing standard tools for characterizing molecular relatedness (Rogers & Hahn, 2010; Bajusz et al., 2015; Riniker & Landrum, 2013; Bemis & Murcko, 1996). Benchmarking efforts such as MoleculeNet, time-aware evaluation, and leakage-conscious split design further show that apparent generalization depends strongly on how training and test compounds are separated (Wu et al., 2018; Sheridan, 2013; Wallach & Heifets, 2018; Yang et al., 2019; Joeres et al., 2025). More recent work, including SAE, DrugOOD, and GOOD, has made this dependence on molecular similarity and shift regime more explicit (Zhang et al., 2025; Ji et al., 2023; Gui et al., 2022; Fooladi et al., 2025). These studies primarily characterize how individual-model performance changes with distribution shift or reduced chemical support; our focus is the complementary question of whether the *relative ordering* of candidate models remains stable across support regimes. Table 1 summarizes this distinction.

This question also connects to model selection under distribution shift. Validation performance may cease to reliably predict target-domain performance once the data distribution changes (Gulrajani & Lopez-Paz, 2021; Krueger et al., 2021; Cha et al., 2021; Ovadia et al., 2019; Rabanser et al., 2019), and model rankings obtained under one evaluation regime need not be preserved under another (Dong et al., 2022; Arpit et al., 2022; Jiang & Teney, 2025; Singh et al., 2026). These findings suggest that leaderboard reliability is itself distribution dependent. In molecular prediction, drug-side chemical support provides a natural continuous axis along which this dependence can be examined.

Table 2: Datasets and candidate-label semantics.

Dataset	Task	Drugs	Counterparts	Positives	Valid structures	Candidate-label semantics
MIFNDRA	ncRNA–drug	121	3,322	7,430	121/121	Unknown as negative
Yamanishi_08	drug–protein	791	989	5,127	791/791	Unknown as negative
MDAD (feature-complete)	drug–microbe	846	159	1,520	846/846	Unknown as negative
ChEMBL37 (direct-human DTI)	drug–protein	1,558	497	2,121	1,558/1,558	Positive–unlabelled

Molecular-association prediction provides a useful setting for this analysis because both evaluation protocols and model inductive biases vary substantially. Prior work has developed cold-start protocols and documented the effects of structural and chemical-series overlap on apparent generalization (Pahikkala et al., 2015; Mayr et al., 2018), while interaction-prediction methods span convolutional, graph-based, bilinear, and transformer architectures (Öztürk et al., 2018; Lee et al., 2019; Nguyen et al., 2021; Li et al., 2020; Chen et al., 2020; Huang et al., 2021; Bai et al., 2023). This diversity makes molecular-association benchmarks well suited to studying whether candidate-model ordering remains informative as training-set chemical support decreases.

3 Chemical-Support-Resolved Evaluation

3.1 Datasets and Task Settings

We evaluate chemical-support-resolved model selection across three molecular-association settings: non-coding RNA (ncRNA)–drug prediction using MIFNDRA (Sui et al., 2026), drug–protein interaction prediction using Yamanishi_08 (Yamanishi et al., 2008), and drug–microbe association prediction using the feature-complete subset of MDAD (Sun et al., 2018). All analyzed drugs have valid canonical structures. Following the native conventions of these benchmarks, unobserved candidate pairs are used as negative candidates for evaluation.

We additionally construct a direct-human DTI subset of ChEMBL37 under positive–unlabelled semantics (Gaulton et al., 2017). Observed direct-human interactions define positives, whereas unobserved drug–protein pairs remain unlabelled and are encoded as zero only for ranking-based evaluation, not as confirmed biological negatives. Each dataset is analyzed separately under its native candidate-label semantics, with drug-side training-set chemical support serving as the common evaluation axis; raw performance values are therefore not compared directly across task families. MDS and MDR are included only as additional RNA–drug robustness settings and are described in Appendix A.1 (Sheng et al., 2026). Table 2 summarizes dataset statistics and label semantics.

3.2 Chemical Support and Evaluation Protocol

Evaluation splits. We use three evaluation protocols, each with an outer test split of approximately 20%. Random Pair performs a stratified 80/20 split over labelled pairs. Drug-Cold instead holds out entire drugs, ensuring that no test drug appears in training. Scaffold-Cold holds out entire Bemis–Murcko scaffold groups while keeping the number of test drugs as close as possible to that of Drug-Cold, thereby eliminating overlap with the training set at both the drug and scaffold levels. For each independent run, the resulting outer partitions are fixed and shared across all models.

Validation data are drawn exclusively from the corresponding outer training partition and follow the same structural constraint as the outer split: pair-stratified for Random Pair, drug-disjoint for Drug-Cold, and scaffold-disjoint for Scaffold-Cold. Hyperparameter selection is completed before evaluation on the outer test set, while the native candidate-label semantics of each benchmark are preserved throughout. The complete partitioning and evaluation protocol is detailed in Appendix A.2.

Operational chemical support. Each drug with a valid canonical Simplified Molecular Input Line Entry System (SMILES) string is represented using a radius-2, 2,048-bit RDKit Morgan fingerprint, and molecular similarity is measured by the Tanimoto coefficient (Rogers & Hahn, 2010; Bajusz et al., 2015; Riniker & Landrum, 2013). For a held-out drug d and the training-drug set $\mathcal{D}_{\text{train}}$, we define its chemical support as

$$S_{\max}(d) = \max_{d' \in \mathcal{D}_{\text{train}}} \text{Tanimoto}\left(\text{Morgan}(d), \text{Morgan}(d')\right).$$

Thus, $S_{\max}$ measures the similarity of a held-out drug to its nearest training-set analogue, with lower values indicating weaker support from the observed training chemistry. We use five pre-specified support intervals: $[0.8, 1.0]$, $[0.6, 0.8)$, $[0.4, 0.6)$, $[0.2, 0.4)$, and $[0, 0.2)$. The chemical-support definition and interval specification are detailed in Appendix A.3.

Bemis–Murcko scaffolds are computed from the same canonical molecular structures. Scaffold membership provides a discrete constraint for split construction, whereas $S_{\max}$ provides a continuous measure of training-set chemical support. Scaffold novelty therefore identifies whether a compound belongs to an unseen scaffold group, while $S_{\max}$ quantifies how closely that compound is represented by the training chemistry.

3.3 Model-Selection Stability and Robustness

Leaderboard stability. Let $\mathcal{M} = \{m_1, \ldots, m_K\}$ denote the candidate models and $A_m(R)$ the performance of model m under regime R. We define the corresponding leaderboard as

$$\pi_R = \mathrm{rank}(\{A_m(R)\}_{m \in \mathcal{M}}),$$

with models ranked by decreasing performance. Random Pair serves as the conventional reference, while support-resolved cohorts are defined within Drug-Cold using $S_{\max}$. Our primary comparison is Random Pair versus the extreme-support regime ($S_{\max} < 0.2$); we additionally compare $S_{\max} \in [0.4, 0.6)$ with $S_{\max} < 0.2$ entirely within Drug-Cold to hold the split protocol fixed. Ranking agreement is measured by Spearman correlation, Kendall's τ, Top-1 agreement, Top-3 overlap, and pairwise order reversals, where Top-3 overlap is $|\mathrm{Top3}_A \cap \mathrm{Top3}_B|/3$.

Models and evaluation metric. We evaluate eight controlled model families: logistic regression (M1), ExtraTrees (M2), random forest (M3), two-tower dot product (M4), two-tower multilayer perceptron (MLP) (M5), bilinear interaction (M6), pair MLP (M7), and k-nearest-neighbour (kNN) transfer (M8). Drug inputs use Morgan fingerprints, while RNA, protein, and microbial counterparts use RNA foundation model (RNA-FM) embeddings, Evolutionary Scale Modeling 2 (ESM-2) embeddings, and fixed National Center for Biotechnology Information (NCBI) taxonomy multi-hot features, respectively (Chen et al., 2022; Lin et al., 2023). Full configurations and training details are provided in Appendix A.4.

Our primary metric is macro drug-level average precision (AP), computed per held-out drug and then averaged across drugs; drugs without observed positive candidates receive AP $= 0$. We additionally report median drug-level AP, micro area under the precision–recall curve (AUPR), area under the receiver operating characteristic curve (AUROC), AP lift, Precision@20, and Hit@20 where compatible with the task semantics. For ChEMBL37, macro AP is interpreted as observed-positive recovery under positive–unlabelled evaluation.

Ranking robustness. Within each run, M1–M8 are ranked by macro AP across Random Pair, Drug-Cold, Scaffold-Cold, and support-resolved Drug-Cold cohorts. A pairwise reversal is considered meaningful only when the macro-AP difference exceeds 0.005 in both regimes; 0.01 is used as a sensitivity threshold. We further assess ranking transitions using whole-drug resampling, covariate adjustment, and $S_{\max}$-matched Scaffold-Cold–Drug-Cold comparisons. The *ex-post oracle selection gap* is defined as the extreme-support AP of the best model minus that of the Random Pair winner.

Sparse-tail and representation controls. Because the $S_{\max} < 0.2$ cohorts contain only two to nine drugs per run, we compare them with 1,000 equal-size samples drawn from the $S_{\max} \in [0.4, 0.6)$ Drug-Cold pool using one-sided Monte Carlo tests, with Holm correction applied separately to the pre-specified statistics. Finite-cohort sensitivity is further examined using whole-drug bootstrap, leave-one-drug-out analysis, and bottom-5%/10% support tails. We also evaluate label-disjoint model selection, recompute $S_{\max}$ with radius-3 Morgan and MACCS-166 fingerprints, and, for MDAD, match moderate- and low-support drugs on positive count, prevalence, molecular weight, and heavy-atom count.

Additional DTI architectures. For Yamanishi_08 and ChEMBL37, we additionally evaluate six representative DTI architectures: CPI-GNN (Tsubaki et al., 2019), MolTrans (Huang et al., 2021),

DrugBAN (Bai et al., 2023), DeepConv-DTI (Lee et al., 2019), TransformerCPI (Chen et al., 2020), and HyperAttentionDTI (Zhao et al., 2022). CPI-GNN, MolTrans, and DrugBAN use fixed pair-level predictions under the common evaluation protocol, whereas DeepConv-DTI, TransformerCPI, and HyperAttentionDTI are trained on the same Random Pair and Drug-Cold partitions using all positives and up to ten sampled negatives per positive, with epochs selected on the corresponding validation split. All six models are evaluated on the same test pairs, Morgan–Tanimoto support values, and macro drug-level AP metric. We further include a task-specific MDAD sensitivity analysis with HMDAKATZ (Zhu et al., 2021), NIRBMMDA (Cheng et al., 2022), and SCSMDA (Tian et al., 2023), using association-derived inputs reconstructed exclusively from outer-training positives. Full configurations and per-regime results are provided in Appendix A.5.

4 RESULTS

4.1 SCAFFOLD-COLD DOES NOT IMPLY UNIFORM CHEMICAL SUPPORT

Scaffold-Cold alters the distribution of chemical support but does not place all held-out drugs in a uniformly low-support regime. In MIFNDRA, Yamanishi_08, and MDAD, 15.3%, 8.2%, and 11.6% of Scaffold-Cold test drugs, respectively, still have $S_{\max} > 0.6$, while 9.7%, 5.5%, and 3.7% exceed 0.7. These compounds are scaffold-disjoint from the training set yet retain relatively close Morgan–Tanimoto analogues, demonstrating that scaffold novelty and training-set chemical support capture distinct properties. Figure 2a summarizes the corresponding support distributions; the complete tabulation is retained in Appendix A.6.

The support shift also varied across tasks. Scaffold-Cold median $S_{\max}$ was consistently lower than Drug-Cold in Yamanishi_08 (0.340–0.352 vs. 0.422–0.476) and MDAD (0.277–0.447 vs. 0.585–0.600), whereas MIFNDRA showed substantial overlap (0.216–0.255 vs. 0.221–0.289). After matching on $S_{\max}$, mean macro-AP differences were near zero in MIFNDRA (-0.0019) and Yamanishi_08 ($+0.0010$); MDAD was less conclusive because only about 30% of candidate drugs could be matched. Full support distributions and matched analyses are reported in Appendix A.6.

4.2 SUPPORT-RESOLVED PERFORMANCE DEGRADATION IS TASK DEPENDENT

Performance declined sharply with decreasing chemical support in Yamanishi_08. Mean macro drug-level AP decreased from 0.484 in the $[0.8, 1.0]$ bin to 0.381, 0.264, 0.094, and 0.023 across progressively lower-support regimes. The median per-drug $S_{\max}$–AP correlation was 0.486 and remained 0.499 after adjustment for positive degree, molecular weight, and heavy-atom count (median $p = 4.62 \times 10^{-11}$), indicating that the association persists after covariate adjustment. Figure 2b summarizes this pattern; complete support-bin and adjusted results are reported in Appendices A.7 and A.8.

ChEMBL37 showed a similar monotonic decline under positive–unlabelled semantics, with observed-positive recovery decreasing from 0.542 to 0.452, 0.280, 0.076, and 0.034 and a mean $S_{\max}$–AP correlation of 0.627. In contrast, MIFNDRA and MDAD showed no comparable monotonic trend, with median correlations of -0.103 and 0.134, respectively. Thus, predictive degradation with decreasing support is pronounced in some settings but remains task and label-regime dependent.

4.3 SUPPORT-CONDITIONED RANKING IS REGIME DEPENDENT

Ranking agreement deteriorated most strongly in the extreme-support regime. Random Pair-to-$S_{\max} < 0.2$ Spearman correlations were -0.302, 0.183, and 0.087 in Yamanishi_08, MDAD, and ChEMBL37, with meaningful pairwise reversal rates of 48.8%, 34.5%, and 34.5%, respectively. Kendall's τ and Top-1 agreement showed the same pattern (Figure 3; Table 3).

The decline in agreement was concentrated toward the low-support tail. Random Pair rankings remained substantially more consistent at $S_{\max} \in [0.4, 0.6)$, with Spearman correlations of 0.841 in Yamanishi_08 and 0.770 in MDAD, but approached zero or became negative at $S_{\max} < 0.2$. The same pattern persisted when the split protocol was held fixed: comparing $S_{\max} \in [0.4, 0.6)$ with $S_{\max} < 0.2$ entirely within Drug-Cold yielded Spearman correlations of -0.373, -0.008,

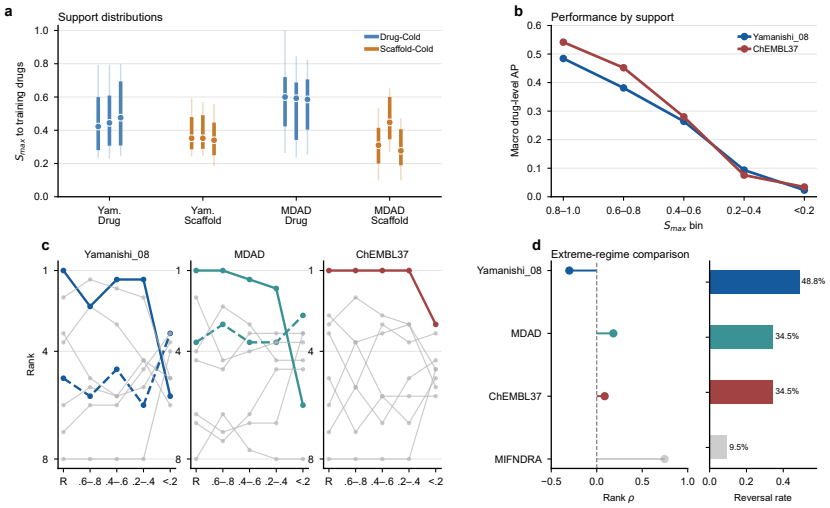


Figure 2: Chemical support separates predictive difficulty from model-order stability. (a) Drug-Cold and Scaffold-Cold support distributions. (b) Mean macro drug-level AP by support bin. (c) Controlled-model rank trajectories. (d) Random Pair-to-extreme rank correlation and meaningful pairwise reversal rate.

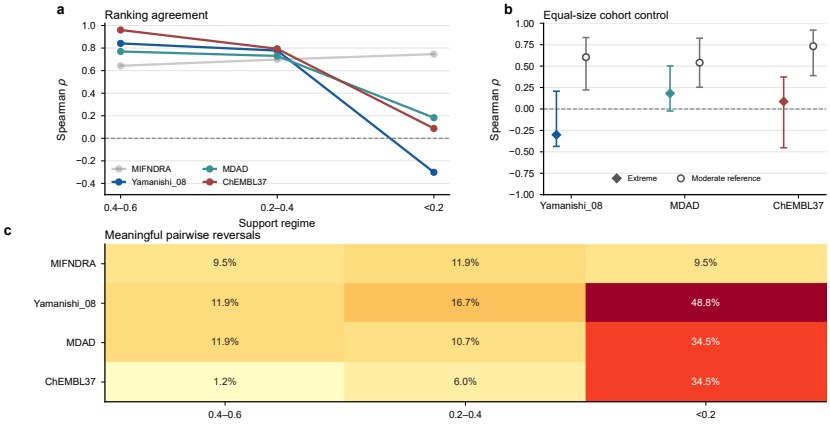


Figure 3: **Leaderboard stability across chemical-support regimes.** (a) Random Pair-to-bin Spearman correlations across support bins; (b) observed extreme-support versus equal-size moderate-support references with 95% intervals; (c) meaningful pairwise reversals for model pairs with pre-specified $\Delta\text{AP} > 0.005$ in both regimes, with stricter-threshold sensitivity in Appendix A.9.

and 0.087 in Yamanishi_08, MDAD, and ChEMBL37, with reversal rates of 48.8%, 44.0%, and 39.3%. Whole-drug resampling and leave-one-drug-out analyses did not identify a single drug as the dominant source of these transitions. Additional run-level, RNA–drug, and ChEMBL37 protocol results are reported in Appendices A.9–A.11.

To assess whether sparse cohort size alone explains the instability, we compared each extreme-support cohort with 1,000 equal-size samples from the $S_{\max} \in [0.4, 0.6)$ Drug-Cold pool. Extreme-support rank correlations were substantially lower in Yamanishi_08 (-0.302 versus 0.604, $p = 0.001$), MDAD (0.183 versus 0.540, $p = 0.011$), and ChEMBL37 (0.087 versus 0.732, $p = 0.001$), with correspondingly higher reversal rates ($p \leq 0.013$). All six pre-specified comparisons remained significant after Holm correction. MIFNDRA was not evaluable because its moderate-support pool was smaller than its extreme-support cohort. These controls indicate that reduced cohort size alone does not explain the loss of ranking agreement (Figure 3b; Appendix A.12).

Table 3: Model-order stability from Random Pair to $S_{\max} < 0.2$, averaged across three runs; meaningful reversals require $\Delta\text{AP} > 0.005$ in both regimes.

Dataset	**Spearman**	Kendall's τ	Top-1 agreement	Top-3 overlap	**Meaningful reversals**
MIFNDRA	0.746	0.571	0/3	0.667	9.5%
Yamanishi_08	−0.302	−0.286	1/3	0.333	48.8%
MDAD	0.183	0.167	0/3	0.333	34.5%
ChEMBL37	0.087	0.024	1/3	0.333	34.5%

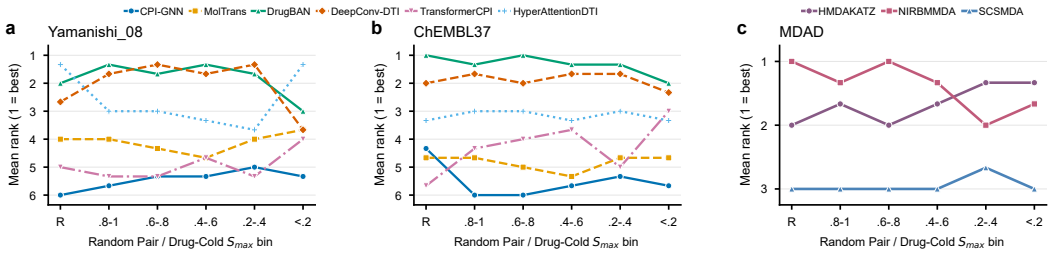


Figure 4: **Ranking stability depends on the evaluated model set.** (a,b) Six-model DTI rank trajectories on Yamanishi_08 and ChEMBL37; (c) HMDAKATZ, NIRBMMDA, and SCSMDA on MDAD. Ranks are computed within each run from macro drug-level AP across the same Random Pair and Drug-Cold regimes, then averaged for display.

4.4 Label-Disjoint Selection in the Low-Support Regime

Ranking transitions can alter which model is preferred in the low-support regime. In the primary Morgan $S_{\max} < 0.2$ cohort, the mean *ex-post oracle gap*—the difference between the best extreme-support model and the Random Pair winner—was 0.038 in Yamanishi_08, 0.130 in MDAD, and 0.032 in ChEMBL37. The corresponding gaps were positive, although drug-level bootstrap uncertainty remained substantial because the extreme-support cohorts contained few drugs (Appendix A.13).

We therefore evaluated model selection using disjoint low-support labels. Each extreme-support cohort was partitioned into balanced subsets, with one subset used to select a model by macro AP and the other used to evaluate that choice. Relative to the Random Pair winner, label-disjoint selection yielded higher held-out macro AP in Yamanishi_08 ($\Delta = 0.010$; 56.4% of partitions), MDAD ($\Delta = 0.059$; 74.3%), and MIFNDRA ($\Delta = 0.029$; 71.9%), but not in ChEMBL37 ($\Delta = -0.016$; 19.8%; Appendix A.13). These results show that support-conditioned ranking changes can translate into different model choices, while the benefit of low-support model selection remains dataset dependent.

4.5 Cross-Task Synthesis and Robustness Analyses

Chemical-support effects differed across tasks. Yamanishi_08 combined monotonic predictive degradation with the clearest low-support ranking transition; MDAD showed weak extreme-support rank agreement without a comparable monotonic AP decline. MIFNDRA retained relatively stable ordering (Random-to-extreme $\rho = 0.746$), whereas ChEMBL37 showed both monotonic degradation and near-zero rank agreement under its distinct positive–unlabelled semantics. Thus, predictive degradation and ranking instability are distinct responses to reduced chemical support. Across six DTI architectures, mean Random-to-extreme Spearman ρ was 0.543 in Yamanishi_08 and 0.467 in ChEMBL37 (Figure 4; Appendix A.5). In the separate three-model MDAD analysis, mean Random-to-extreme ρ was 0.667, with an 11.1% meaningful pairwise reversal rate. Morgan-r3 preserved the Yamanishi_08 transition and MDAD/ChEMBL37 reversal patterns, and agreed strongly with Morgan-r2 ($\rho = 0.978$–0.987); MACCS agreement was lower ($\rho = 0.601$–0.725) and attenuated MDAD and ChEMBL37 transitions. Precision@20 and Hit@20 retained near-zero Random-to-extreme agreement in Yamanishi_08 and ChEMBL37, while MDAD correlations increased to 0.431 and 0.366. After matching MDAD drugs on positive count, prevalence, molecular weight, and

Table 4: Minimum reporting set for chemical-support-resolved evaluation.

Component	What should be reported	Purpose
Split definition	Random Pair, Drug-Cold, Scaffold-Cold, and overlap constraints	States what is prevented
Chemical support	Representation, similarity, $S_{\max}$ distribution, and bin sizes	States the support regime
Absolute performance	Per-drug, support-resolved score	Measures difficulty
Model-order stability	Rank correlation, Top-1 agreement, meaningful pairwise reversals	Assesses leaderboard stability
Uncertainty / boundary	Tail size, equal-size moderate-support reference or bootstrap, alternative representation	Bounds finite-cohort and representation dependence

heavy-atom count, agreement remained low ($\rho = 0.087$), with 42.9% meaningful pairwise reversals and no Top-1 agreement. Overall, results depended on the task, candidate model set, representation, and metric; complete sensitivity results and the evidence summary are in Appendices A.14 and A.15.

5 Discussion

5.1 Interpreting Support-Resolved Evaluation

Chemical-support-resolved evaluation complements conventional molecular split protocols by separating discrete overlap constraints from continuous training-set support. Drug-Cold and Scaffold-Cold control specific forms of train–test overlap, whereas $S_{\max}$ captures how strongly each held-out drug is represented by the training chemistry.

Predictive degradation and model-order stability need not vary together. Lower support may reduce predictive performance while preserving model ordering, or alter model rankings without a comparable monotonic decline in performance. Yamanishi_08 and ChEMBL37 exhibit both support-dependent degradation and low-support ranking transitions, whereas MDAD shows weak low-support rank agreement without a comparable AP trajectory. Similar behavior across additional DTI architectures indicates that the magnitude of ranking change also depends on the candidate model set.

Model rankings observed in one support regime therefore need not persist in another. Benchmark leaderboards should be interpreted relative to the chemical-support regime in which model selection is applied. These associations are descriptive rather than causal; reduced chemical support alone does not establish the cause of a ranking transition.

5.2 Practical Reporting Recommendations

In practice, support-resolved evaluation is most informative when conventional split definitions are reported together with training-set support distributions, support-resolved predictive performance, model-order stability, and finite-cohort uncertainty. Because support depends on the molecular representation and similarity definition, threshold-dependent findings should also be checked under reasonable alternative representations and similarity measures. Table 4 summarizes these practical reporting elements.

Limitations. Chemical support is representation dependent: radius-3 Morgan fingerprints yield similar estimates, whereas alternatives such as MACCS can alter the composition of low-support cohorts. Conclusions about ranking stability also remain conditional on the candidate models, task domains, and label semantics considered here, and our analysis focuses on drug-side chemical support. Future work should examine broader model and task families, prospective or temporally separated evaluation settings, alternative notions of chemical support, and joint drug- and counterpart-side extrapolation.

Conclusion. Molecular benchmarks should distinguish split membership from chemical support and predictive performance from model-order stability. We introduced chemical-support-resolved evaluation to assess both across continuous test-to-training support. Across molecular-association tasks, Scaffold-Cold spanned heterogeneous support regimes, with performance degradation and ranking shifts under low support that depended on model architecture, representation, and support definition. Reporting support distributions, tail performance, and ranking stability alongside standard split-level metrics can therefore better assess whether benchmark leaderboards remain informative for model selection.

AI USE STATEMENT

We used language models to assist with language editing, code organization, and provenance checking. All experimental design, implementation decisions, result interpretation, and final verification were performed by the authors. Language-model outputs were manually reviewed for technical accuracy and alignment with the paper's objectives.

ETHICS STATEMENT

All authors adhere to the ICLR Code of Ethics. This study investigates evaluation and model-selection reliability in molecular association prediction using existing benchmark datasets, including ncRNA–drug, drug–protein, and drug–microbe interaction data. No new human participants, animal experiments, clinical interventions, or personally identifiable information were collected in this work. The analyses rely on previously released datasets and retain their original label semantics and usage constraints. The proposed chemical-support-resolved evaluation framework is intended as a computational research tool for assessing model generalization and selection under molecular extrapolation. Any downstream use for biological or drug-discovery decision making should be accompanied by appropriate experimental validation and domain-specific assessment.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, chemical-support definition, data-splitting protocols, model definitions, training and validation procedures, evaluation metrics, and statistical analyses in the main text and Appendix. All primary experiments are repeated across three independent runs using pre-specified Random Pair, Drug-Cold, and Scaffold-Cold partitions, fixed molecular representations, and consistent aggregation rules. The Appendix further documents dataset provenance, model configurations and hyperparameter-search spaces, run-level results, support-bin statistics, uncertainty analyses, representation sensitivity, matching analyses, and additional architecture evaluations. The accompanying materials provide the scripts and configuration files required to reproduce the reported analyses.

REFERENCES

Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35:8265–8277, 2022.

Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20, 2015.

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, Irwin King, and Yu Li. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022. doi: 10.48550/arXiv.2204.00300.

Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 2020.

Xiaolong Cheng, Jia Qu, Shuangbao Song, and Zekang Bian. Neighborhood-based inference and restricted boltzmann machine for microbe and drug associations prediction. *PeerJ*, 10:e13848, 2022.

Qishi Dong, Awais Muhammad, Fengwei Zhou, Chuanlong Xie, Tianyang Hu, Yongxin Yang, Sung-Ho Bae, and Zhenguo Li. Zood: Exploiting model zoo for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:31583–31598, 2022.

Hosein Fooladi, Thi Ngoc Lan Vu, Miriam Mathea, and Johannes Kirchmair. Evaluating machine learning models for molecular property prediction: performance and robustness on out-of-distribution data. *Journal of Chemical Information and Modeling*, 65(19):9871, 2025.

Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8023–8031, 2023.

Liangze Jiang and Damien Teney. OOD-chameleon: Is algorithm selection for OOD generalization learnable? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27624–27648. PMLR, 2025.

Roman Joeres, David B Blumenthal, and Olga V Kalinina. Data splitting to avoid information leakage with datasail. *Nature communications*, 16(1):3337, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6): e1007129, 2019.

Shuya Li, Fangping Wan, Hantao Shu, Tao Jiang, Dan Zhao, and Jianyang Zeng. Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell systems*, 10(4):308–322, 2020.

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

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical science*, 9(24):5441–5451, 2018.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pp. 7721–7735. PMLR, 2021.

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16(2):325–337, 2015.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Sereina Riniker and Gregory A Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics*, 5(1):26, 2013.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Nan Sheng, Yunzhi Liu, Ling Gao, Wenju Hou, Lan Huang, and Yan Wang. A pre-trained language model-based cross-modal fusion framework for predicting miRNA-drug resistance and sensitivity associations. *PLOS Computational Biology*, 22(2):e1013968, 2026.

Robert P Sheridan. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of chemical information and modeling*, 53(4):783–790, 2013.

Robert P Sheridan, Bradley P Feuston, Vladimir N Maiorov, and Simon K Kearsley. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of chemical information and computer sciences*, 44(6):1912–1928, 2004.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel Dsouza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. The leaderboard illusion. *Advances in neural information processing systems*, 38, 2026.

Jianan Sui, Weirong Cui, Xiaojie Jin, Hongliang Duan, and Jingjing Guo. MIFNDRA: an innovative knowledge-enhanced multimodal fusion and graph learning framework for predicting drug resistance-related ncrnas. *Briefings in Bioinformatics*, 27(3):bbag322, 2026.

Ya-Zhou Sun, De-Hong Zhang, Shu-Bin Cai, Zhong Ming, Jian-Qiang Li, and Xing Chen. MDAD: A special resource for microbe–drug associations. *Frontiers in cellular and infection microbiology*, 8:424, 2018.

Zhen Tian, Yue Yu, Haichuan Fang, Weixin Xie, and Maozu Guo. Predicting microbe–drug associations with structure-enhanced contrastive learning and self-paced negative sampling strategy. *Briefings in bioinformatics*, 24(2):bbac634, 2023.

Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 2019.

Izhar Wallach and Abraham Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of chemical information and modeling*, 58(5):916–932, 2018.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240, 2008.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370, 2019.

Chenbin Zhang, Zhiqiang Hu, Jiang Chuchu, Wen Chen, Jie Xu, and Shaoting Zhang. Rethinking the generalization of drug target affinity prediction algorithms via similarity aware evaluation. In *International Conference on Learning Representations*, volume 2025, pp. 45891–45917, 2025.

Qichang Zhao, Haochen Zhao, Kai Zheng, and Jianxin Wang. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. *Bioinformatics*, 38(3):655–662, 2022.

Lingzhi Zhu, Guihua Duan, Cheng Yan, and Jianxin Wang. Prediction of microbe-drug associations based on chemical structures and the katz measure. *Current Bioinformatics*, 16(6):807–819, 2021.

A Supplementary Methods and Results

A.1 Dataset Construction and Provenance

We first define the final set of eligible drugs, counterparts, and associations before constructing any evaluation split. For each data source, drug and counterpart identifiers are reconciled with the corresponding association records, canonical SMILES are mapped to drug entities, and records lacking the molecular or counterpart information required by the evaluation protocol are excluded from the final analysis set while remaining traceable in the provenance record. All subsequent Random Pair, Drug-Cold, and Scaffold-Cold partitions are therefore constructed from the same valid-SMILES, feature-complete data.

The primary benchmarks retain their original candidate-label conventions after preprocessing. MDS and MDR are included only as additional RNA–drug robustness settings and follow the same preprocessing family. ChEMBL37 is analyzed separately in a direct-human drug–protein evaluation under positive–unlabelled semantics: observed interactions define positives, whereas unobserved drug–protein candidates remain unlabelled and are used only as the candidate complement for ranking-based evaluation. Table 5 summarizes the source versions, raw entity and interaction counts, candidate semantics, and the filtering boundary used to obtain each final analysis universe.

Table 5: Dataset provenance and preprocessing. The table summarizes the source records, candidate-label semantics, and filtering criteria used to obtain the final analysis sets. All reported splits are constructed from records with valid canonical SMILES and the features required by the corresponding task.

Dataset	Source / version	Raw entities	Raw observed edges	Fixed preprocessing and final universe
MIFNDRA	MIFNDRA release	121 drugs; 3,322 RNA	7,430 positives; 394,532 complement candidates	valid canonical SMILES; unknown-as-negative; 121/3,322/7,430
Yamanishi_08	Combined Yamanishi_08 benchmark	791 drugs; 989 proteins	5,127 positives; 777,172 complement candidates	valid canonical SMILES; unknown-as-negative; 791/989/5,127
MDAD	Feature-complete MDAD audit universe	846 drugs; 159 microbes	1,520 positives; 132,994 complement candidates	valid canonical SMILES and taxonomy coverage; unknown-as-negative; 846/159/1,520
MDS	Author-provided MDS matrix	140 drugs; 1,252 RNA	5,054 positives; 170,226 explicit negatives	valid canonical SMILES; explicit-negative semantics; 140/1,252/5,054
MDR	Author-provided MDR matrix	105 drugs; 1,317 RNA	5,411 positives; 132,874 explicit negatives	valid canonical SMILES; explicit-negative semantics; 105/1,317/5,411
ChEMBL37	Direct-human ChEMBL37 extract	1,558 drugs; 497 proteins	2,121 observed positives; complement unlabelled	valid canonical SMILES; positive–unlabelled ranking; 1,558/497/2,121

A.2 Pre-specified Partitions and Evaluation Metrics

For MIFNDRA, Yamanishi_08, MDAD, MDS, and MDR, evaluation is repeated over three independent runs using partitions fixed before model fitting. Random Pair holds out labelled pairs while allowing the participating entities to remain represented in training. Drug-Cold instead holds out entire drugs, eliminating train–test drug overlap, whereas Scaffold-Cold holds out entire Bemis–Murcko scaffold groups and therefore removes scaffold overlap between training and test compounds. Validation data are drawn exclusively from the corresponding training partition, and hyperparameters are selected before evaluation on the held-out test set. Within each run, the same outer partitions are shared across all models and downstream analyses.

Our primary evaluation metric is macro drug-level AP. Within each test partition, AP is computed separately for every test drug over its full set of candidate counterparts and then averaged across drugs, giving each drug equal weight in the aggregate score. Drugs with no positive candidate are assigned $\mathrm{AP} = 0$. We additionally report micro AUPR, AUROC, AP lift, Precision@20, and Hit@20 where these metrics are compatible with the corresponding candidate-label semantics. Random Pair primarily characterizes interpolation performance, whereas Drug-Cold and its support-resolved subsets characterize drug-side extrapolation at progressively weaker levels of training-set chemical support. Throughout the analysis, the pre-specified $S_{\max} < 0.2$ subset is referred to as the *extreme-support regime*.

A.3 Chemical Support Definition

All drug-side chemical comparisons are based on a common structural representation. Each compound is encoded as a radius-2, 2048-bit RDKit Morgan fingerprint, and pairwise chemical similarity is measured using the Tanimoto coefficient. Bemis–Murcko scaffolds are derived from the same standardized molecular records so that the discrete scaffold-based split definition and the continuous support measure are computed from a consistent chemical representation.

For a held-out drug d, we define its training-set chemical support as the similarity to its nearest training compound,

$$S_{\max}(d) = \max_{d' \in \mathcal{D}_{\text{train}}} \text{Tanimoto}(\text{Morgan}(d), \text{Morgan}(d')) .$$

Larger values therefore indicate that the test drug has a closer structural analogue in the training set, whereas smaller values correspond to weaker local support from the observed training chemistry.

All support-resolved analyses use the same pre-specified intervals: $[0.8, 1.0]$, $[0.6, 0.8)$, $[0.4, 0.6)$, $[0.2, 0.4)$, and $[0, 0.2)$. These bins are applied within the Drug-Cold test sets to stratify held-out drugs by their level of training-set chemical support. The corresponding cohort sizes are reported in the subsequent support-bin analyses.

A.4 Controlled Model Configurations

The controlled M1–M8 model set is evaluated under a common feature-construction and model-selection protocol, while retaining model-specific architectures, search spaces, and optimization budgets. All preprocessing and hyperparameter selection are confined to the corresponding outer training partition. Table 6 summarizes the model architectures, candidate hyperparameters, optimization settings, and validation criteria used throughout the evaluation.

Representations and preprocessing. Drug inputs use the Morgan fingerprints defined in Appendix A.3. RNA counterparts are represented using the released 640-dimensional RNA-FM embeddings after uppercase T-to-U normalization; long sequences are processed with a 1,024-token context and a 1,022-token stride, followed by mean pooling across chunks. Protein counterparts use frozen ESM-2 `t6_8M_UR50D` embeddings, mean-pooled over residues to obtain 320-dimensional sequence representations. Microbial counterparts are encoded using NCBI taxonomy-rank multi-hot features derived from the audited MDAD taxonomy records. Continuous counterpart features are standardized using statistics estimated from outer-training entities only.

For the classical models, drug and counterpart representations are independently mapped to deterministic 128-dimensional Gaussian projections before constructing the joint feature vector $[a, b, a \odot b, |a - b|]$. Neural models instead operate directly on the corresponding standardized side representations. All learned transformations are fit within the outer training partition, and model selection is performed using only its associated validation split before evaluation on the held-out test data.

A.5 Support-Conditioned Ranking Stability Across DTI Architectures

To examine whether support-conditioned changes in model ordering persist beyond the controlled M1–M8 model set, we extend the analysis to six representative DTI architectures: CPI-GNN, MolTrans, DrugBAN, DeepConv-DTI, TransformerCPI, and HyperAttentionDTI. These models span graph-based, substructure-aware, bilinear-attention, convolutional neural network (CNN)-based, and attention-based interaction mechanisms, providing a complementary architecture-level test of ranking stability. The evaluation is conducted on Yamanishi_08 and ChEMBL37, the two drug–protein settings for which the same candidate-model comparison can be applied under a common protocol.

Within each dataset and independent run, all six architectures are evaluated on identical outer partitions, test pairs, candidate labels, $S_{\max}$ assignments, and pre-specified chemical-support intervals. Model selection is restricted to validation data derived from the corresponding training partition, so support-resolved test rankings are formed only after the training configuration has been fixed. Architecture-specific optimization, training budgets, and validation-based selection rules are reported in Table 7.

We evaluate each architecture under Random Pair, overall Drug-Cold, and progressively lower-support subsets of the Drug-Cold test set. This provides a common basis for examining whether changes in predictive performance are accompanied by changes in model ordering as training-set chemical support decreases. Table 8 reports the corresponding macro drug-level AP profiles across evaluation regimes.

Because aggregate performance alone does not determine whether the same model would be selected, we additionally recover the model ordering within each independent run and support regime.

Table 6: Controlled-model configurations and model-selection protocol. Architectures and hyperparameter search spaces are held fixed across datasets, and final configurations are selected by validation macro drug-level AP within the corresponding training partition. All learned models use the same deterministic training sample containing all positives and up to 10× sampled negatives.

ID	Model	Model specification	Hyperparameters / training
M1	Linear logistic	L_2-regularized logistic regression on the 512-dimensional projected pair representation $[a, b, a \odot b, \lvert a - b \rvert]$	$C \in \{0.1, 1.0\}$; class-balanced L-BFGS; maximum 300 iterations; tolerance 10^{-3}.
M2	ExtraTrees	100 randomized trees; maximum depth 16; square-root feature subsampling	Minimum leaf size $\in \{1, 5\}$; class-balanced weighting; deterministic initialization.
M3	Random forest	100 trees; maximum depth 16; square-root feature subsampling	Minimum leaf size $\in \{1, 5\}$; balanced-subsample weighting; deterministic initialization.
M4	Two-tower dot	Separate drug and counterpart encoders to 128 dimensions with Gaussian error linear unit (GELU) and dropout 0.1; scaled dot-product interaction	AdamW; learning rate $\in \{10^{-3}, 3 \times 10^{-4}\}$; weight decay 10^{-4}; weighted binary cross-entropy (BCE); batch size 1,024; maximum 25 epochs; validation macro-AP early stopping (patience 4).
M5	Two-tower MLP	Same 128-dimensional encoders as M4; interaction head $512 \to 128 \to 64 \to 1$ on $[a, b, a \odot b, \lvert a - b \rvert]$	Same optimization and early-stopping protocol as M4.
M6	Bilinear interaction	Same 128-dimensional encoders as M4; rank-16 bilinear interaction with linear side terms	Same optimization and early-stopping protocol as M4.
M7	Pair MLP	Same 128-dimensional encoders as M4; concatenation head $256 \to 128 \to 64 \to 1$	Same optimization and early-stopping protocol as M4.
M8	kNN transfer	Label transfer from cosine-nearest drug and counterpart neighbours, averaged when both are available	$k \in \{5, 10\}$; no learned parameters or gradient-based optimization.

Table 7: Configuration and selection scope for the six-model DTI architecture set. DeepConv-DTI, TransformerCPI, and HyperAttentionDTI use training-partition validation for epoch selection; test labels are accessed after the selected epoch is frozen.

Model	Defining interaction design	Training budget	Selection rule
CPI-GNN	graph + sequence	Frozen pair-level predictions; Adam 10^{-3}; batch 64; max 100	Validation-selected frozen predictions
MolTrans	substructure transformer	Frozen pair-level predictions; Adam 10^{-4}; batch 16; max 13	Validation-selected frozen predictions
DrugBAN	bilinear attention	Frozen pair-level predictions; Adam 5×10^{-5}; batch 64; max 100	Validation-selected frozen predictions
DeepConv-DTI	Morgan MLP + protein CNN	Adam, 10^{-4}; batch 32; max 15	Validation AP; patience 4
TransformerCPI	molecular graph + protein CNN/attention	Adam, 10^{-3}; weight decay 10^{-4}; max 40	Validation AP; patience 4
HyperAttentionDTI	SMILES/protein CNN + hyperattention	Adam, 5×10^{-5}; weight decay 10^{-4}; max 200	Validation AP; patience 10

Table 8: Six-model DTI macro drug-level AP across chemical-support regimes. Values are means over three independent runs. Ranks are shown for Random Pair and the $S_{max} < 0.2$ cohort.

Dataset	Model	Random	Drug-Cold	.8–1	.6–.8	.4–.6	.2–.4	<.2	Rank R / <.2
Yamanishi_08	CPI-GNN	0.010	0.051	0.066	0.083	0.039	0.035	0.015	6 / 6
Yamanishi_08	MolTrans	0.134	0.116	0.187	0.170	0.129	0.068	0.025	4 / 4
Yamanishi_08	DrugBAN	0.274	0.280	0.543	0.431	0.303	0.146	0.037	2 / 2
Yamanishi_08	DeepConv-DTI	0.260	0.291	0.567	0.452	0.300	0.163	0.030	3 / 3
Yamanishi_08	TransformerCPI	0.063	0.074	0.091	0.083	0.103	0.049	0.020	5 / 5
Yamanishi_08	HyperAttentionDTI	0.279	0.177	0.304	0.297	0.196	0.082	0.039	1 / 1
ChEMBL37	CPI-GNN	0.039	0.028	0.036	0.039	0.037	0.022	0.010	6 / 6
ChEMBL37	MolTrans	0.050	0.066	0.138	0.129	0.080	0.034	0.015	4 / 5
ChEMBL37	DrugBAN	0.110	0.270	0.588	0.572	0.367	0.112	0.095	1 / 1
ChEMBL37	DeepConv-DTI	0.107	0.247	0.548	0.500	0.347	0.105	0.035	2 / 2
ChEMBL37	TransformerCPI	0.043	0.078	0.165	0.170	0.101	0.035	0.023	5 / 3
ChEMBL37	HyperAttentionDTI	0.056	0.129	0.277	0.322	0.146	0.055	0.018	3 / 4

Table 9 reports these run-level rankings, making it possible to distinguish gradual performance degradation from explicit rank transitions between candidate architectures.

The additional architecture set exhibits support-conditioned ordering changes, although the effect is weaker than in the controlled M1–M8 comparison. On Yamanishi_08, the Random Pair and $S_{\max} < 0.2$ rankings yield a mean Spearman correlation of $\rho = 0.543$ and Kendall's $\tau = 0.511$, with Top-1 agreement in one of three runs and a 13.3% meaningful pairwise reversal rate. The run-mean ordering also varies across evaluation regimes: HyperAttentionDTI ranks first under Random Pair and in the extreme-support regime, whereas DeepConv-DTI ranks first under overall Drug-Cold. ChEMBL37 shows a similar but somewhat more stable pattern, with mean Spearman $\rho = 0.467$, Kendall's $\tau = 0.378$, Top-1 agreement in two of three runs, and 13.3% meaningful pairwise reversals. DrugBAN remains the highest-ranked model in the corresponding run-mean summaries. Table 10 summarizes these architecture-level stability statistics across progressively lower-support regimes.

These results show that support-conditioned model ordering is not restricted to the controlled model family, while also indicating that the magnitude of the effect depends on the candidate architectures being compared. The six-model DTI comparison retains non-zero ranking transitions but exhibits substantially greater agreement than the controlled M1–M8 analysis. Leaderboard stability should

Table 9: Run-level ranks for the six-model DTI architecture set. Rank 1 is best; all ranks use macro drug-level AP.

Dataset	Run	Model	Random	.8–1	.6–.8	.4–.6	.2–.4	<.2
Yamanishi_08	Run 1	CPI-GNN	6	6	6	6	6	6
Yamanishi_08	Run 1	MolTrans	4	4	4	4	4	4
Yamanishi_08	Run 1	DrugBAN	2	1	2	1	2	3
Yamanishi_08	Run 1	DeepConv-DTI	3	2	1	2	1	2
Yamanishi_08	Run 1	TransformerCPI	5	5	5	5	5	5
Yamanishi_08	Run 1	HyperAttentionDTI	1	3	3	3	3	1
Yamanishi_08	Run 2	CPI-GNN	6	6	6	6	6	6
Yamanishi_08	Run 2	MolTrans	4	4	4	5	3	4
Yamanishi_08	Run 2	DrugBAN	3	2	2	2	2	1
Yamanishi_08	Run 2	DeepConv-DTI	2	1	1	1	1	3
Yamanishi_08	Run 2	TransformerCPI	5	5	5	3	5	5
Yamanishi_08	Run 2	HyperAttentionDTI	1	3	3	4	4	2
Yamanishi_08	Run 3	CPI-GNN	6	5	4	4	3	4
Yamanishi_08	Run 3	MolTrans	4	4	5	5	5	3
Yamanishi_08	Run 3	DrugBAN	1	1	1	1	1	5
Yamanishi_08	Run 3	DeepConv-DTI	3	2	2	2	2	6
Yamanishi_08	Run 3	TransformerCPI	5	6	6	6	6	2
Yamanishi_08	Run 3	HyperAttentionDTI	2	3	3	3	4	1
ChEMBL37	Run 1	CPI-GNN	4	6	6	5	4	5
ChEMBL37	Run 1	MolTrans	5	5	5	6	5	6
ChEMBL37	Run 1	DrugBAN	1	1	1	1	1	1
ChEMBL37	Run 1	DeepConv-DTI	2	2	2	2	2	2
ChEMBL37	Run 1	TransformerCPI	6	4	4	3	6	4
ChEMBL37	Run 1	HyperAttentionDTI	3	3	3	4	3	3
ChEMBL37	Run 2	CPI-GNN	6	6	6	6	6	6
ChEMBL37	Run 2	MolTrans	4	4	5	5	4	5
ChEMBL37	Run 2	DrugBAN	1	2	1	2	2	4
ChEMBL37	Run 2	DeepConv-DTI	2	1	2	1	1	1
ChEMBL37	Run 2	TransformerCPI	5	5	4	4	5	3
ChEMBL37	Run 2	HyperAttentionDTI	3	3	3	3	3	2
ChEMBL37	Run 3	CPI-GNN	3	6	6	6	6	6
ChEMBL37	Run 3	MolTrans	5	5	5	5	5	3
ChEMBL37	Run 3	DrugBAN	1	1	1	1	1	1
ChEMBL37	Run 3	DeepConv-DTI	2	2	2	2	2	4
ChEMBL37	Run 3	TransformerCPI	6	4	4	4	4	2
ChEMBL37	Run 3	HyperAttentionDTI	4	3	3	3	3	5

Table 10: Six-model DTI ranking stability from Random Pair to progressively lower chemical support. Values are run means. Meaningful pairwise reversals require $|\Delta\mathrm{AP}| > 0.005$ in both regimes.

Dataset	Target regime	Spearman ρ	Kendall τ	Top-1	Top-3 overlap	Reversal rate
Yamanishi_08	0.4–0.6	0.714	0.600	0.333	0.889	0.178
Yamanishi_08	0.2–0.4	0.657	0.556	0.333	0.778	0.222
Yamanishi_08	¡0.2	0.543	0.511	0.333	0.778	0.133
ChEMBL37	0.4–0.6	0.714	0.600	0.667	0.778	0.133
ChEMBL37	0.2–0.4	0.848	0.778	0.667	0.889	0.089
ChEMBL37	¡0.2	0.467	0.378	0.667	0.667	0.133

therefore be interpreted jointly with the evaluated model set rather than as a property of chemical support alone.

We further examine this boundary in MDAD using the task-specific methods HMDAKATZ, NIRB-MMDA, and SCSMDA. Because these methods rely on association-derived inputs, all such inputs are reconstructed exclusively from positive associations in the corresponding outer training partition, preventing test associations from entering model construction. Across the three independent runs, the Random Pair-to-$S_{\max} < 0.2$ comparison yields mean Spearman $\rho = 0.667$, Kendall's $\tau = 0.556$, Top-1 agreement in one of three runs, and an 11.1% meaningful pairwise reversal rate under the 0.005 macro-AP threshold. The MDAD analysis provides complementary evidence on ranking stability under a distinct task-specific model family.

A.6 Scaffold-Support Analyses

Support distributions under Scaffold-Cold. Scaffold-Cold removes Bemis–Murcko scaffold overlap between training and test compounds, but the resulting test drugs need not occupy a uni-

Table 11: Scaffold-Cold induces heterogeneous chemical support. Median values are reported for the three independent runs in order. The matched difference is Scaffold-Cold minus Drug-Cold macro drug-level AP after matching on $S_{\max}$.

Dataset	Drug-Cold median $S_{\max}$ (runs 1–3)	Scaffold-Cold median $S_{\max}$ (runs 1–3)	$S_{\max} > 0.6$	$S_{\max} > 0.7$	Matched ΔAP
MIFNDRA	0.221 / 0.289 / 0.256	0.255 / 0.216 / 0.225	15.3%	9.7%	-0.0019
Yamanishi_08	0.422 / 0.445 / 0.476	0.352 / 0.352 / 0.340	8.2%	5.5%	$+0.0010$
MDAD	0.600 / 0.593 / 0.585	0.310 / 0.447 / 0.277	11.6%	3.7%	$-0.0520^{\dagger}$

† MDAD Scaffold-Cold-to-Drug-Cold matching retained about 30% of candidate drugs at the primary matching threshold and varied across matching-threshold sensitivity analyses.

formly low-support region of chemical space. Across the three primary datasets, the effect of scaffold exclusion on $S_{\max}$ is heterogeneous. MIFNDRA shows substantial overlap between the Drug-Cold and Scaffold-Cold support distributions, whereas Yamanishi_08 and MDAD exhibit more consistent shifts toward lower training-set similarity. Table 12 reports the run-level quantiles under both protocols, showing that scaffold exclusion changes the distribution of chemical support in a dataset-dependent manner rather than defining a single extrapolation regime. Table 11 summarizes the corresponding Scaffold-Cold support statistics.

Table 12: Chemical-support distributions under Drug-Cold and Scaffold-Cold. Run-level quantiles of $S_{\max}$ characterize the training-set chemical support of held-out drugs under the two evaluation protocols. The comparison shows how scaffold exclusion changes the support distribution across the three primary datasets.

Dataset	Run	Split	p10	p25	Median	p75	p90
MDAD	Run 1	Drug-Cold	0.263	0.423	0.600	0.720	1.000
MDAD	Run 1	Scaffold-Cold	0.098	0.200	0.310	0.415	0.533
MDAD	Run 2	Drug-Cold	0.235	0.342	0.593	0.688	0.845
MDAD	Run 2	Scaffold-Cold	0.270	0.345	0.447	0.600	0.651
MDAD	Run 3	Drug-Cold	0.254	0.404	0.585	0.706	0.826
MDAD	Run 3	Scaffold-Cold	0.098	0.189	0.277	0.406	0.470
MIFNDRA	Run 1	Drug-Cold	0.167	0.190	0.221	0.319	0.578
MIFNDRA	Run 1	Scaffold-Cold	0.195	0.210	0.255	0.421	0.722
MIFNDRA	Run 2	Drug-Cold	0.188	0.208	0.289	0.525	0.733
MIFNDRA	Run 2	Scaffold-Cold	0.098	0.183	0.216	0.555	0.741
MIFNDRA	Run 3	Drug-Cold	0.166	0.199	0.256	0.451	0.616
MIFNDRA	Run 3	Scaffold-Cold	0.194	0.203	0.225	0.320	0.450
Yamanishi_08	Run 1	Drug-Cold	0.234	0.281	0.422	0.599	0.793
Yamanishi_08	Run 1	Scaffold-Cold	0.242	0.286	0.352	0.479	0.593
Yamanishi_08	Run 2	Drug-Cold	0.231	0.307	0.445	0.609	0.794
Yamanishi_08	Run 2	Scaffold-Cold	0.245	0.289	0.352	0.491	0.570
Yamanishi_08	Run 3	Drug-Cold	0.246	0.308	0.476	0.694	0.800
Yamanishi_08	Run 3	Scaffold-Cold	0.185	0.250	0.340	0.446	0.560

Matched comparison at comparable chemical support. To separate scaffold novelty from differences in training-set chemical support, we additionally compare Scaffold-Cold and Drug-Cold test drugs after one-to-one matching on $S_{\max}$. Matching is performed using pre-specified $S_{\max}$ difference thresholds so that paired compounds have comparable nearest-neighbour similarity to the corresponding training set. After matching, the mean macro drug-level AP difference between Scaffold-Cold and Drug-Cold is close to zero in MIFNDRA and Yamanishi_08. MDAD shows a larger nominal difference, but only about 30% of candidate drugs are retained under the primary matching threshold, making the estimate more sensitive to the matching specification. Table 13

reports the retained coverage, residual $S_{\max}$ mismatch, and matched macro-AP differences across different matching thresholds.

Table 13: Support-matched comparison of Scaffold-Cold and Drug-Cold. Scaffold-Cold and Drug-Cold test drugs are matched one-to-one using pre-specified $S_{\max}$ difference thresholds. The table reports the fraction of drugs retained after matching, the remaining difference in chemical support, and the matched macro drug-level AP difference, defined as Scaffold-Cold minus Drug-Cold.

Dataset	Caliper	Retained	Mean $\lvert\Delta S_{\max}\rvert$	Scaffold minus Drug-Cold AP
MDAD	0.025	21.5%	0.010	-0.0263
MDAD	0.050	30.0%	0.019	-0.0520
MDAD	0.100	40.6%	0.034	-0.0035
MIFNDRA	0.025	45.8%	0.007	-0.0013
MIFNDRA	0.050	62.5%	0.015	-0.0019
MIFNDRA	0.100	75.0%	0.024	-0.0007
Yamanishi_08	0.025	41.1%	0.009	+0.0147
Yamanishi_08	0.050	54.0%	0.016	+0.0010
Yamanishi_08	0.100	69.4%	0.029	-0.0126

A.7 Support-Resolved Cohort Sizes and Model Performance

Support-resolved cohorts vary in both size and class composition across datasets. This variation is most pronounced in MIFNDRA, where several intermediate-support bins contain only a small number of test drugs. Table 14 reports the mean numbers of drugs, observed positives, candidate pairs, and positive prevalence within each Drug-Cold support bin across the three independent runs. The descriptive counts include all Drug-Cold test drugs available for per-model evaluation.

Table 14: Drug-Cold cohort composition across chemical-support regimes. The table reports the mean numbers of test drugs, observed positives, candidate pairs, and positive prevalence within each support interval across three independent runs.

Dataset	$S_{\max}$ bin	Drugs	Positives	Pairs	Prevalence
MDAD	0.8–1.0	12.2	25.2	1934	1.22%
MDAD	0.6–0.8	40.5	84.8	6440	1.36%
MDAD	0.4–0.6	44.3	76.5	7049	1.08%
MDAD	0.2–0.4	52.8	91.5	8400	1.06%
MDAD	0.0–0.2	19.2	26.0	3048	0.79%
MIFNDRA	0.8–1.0	1.0	23.0	3322	0.54%
MIFNDRA	0.6–0.8	2.7	64.3	8859	0.48%
MIFNDRA	0.4–0.6	2.8	298.7	9412	3.03%
MIFNDRA	0.2–0.4	11.0	694.3	36542	1.62%
MIFNDRA	0.0–0.2	6.5	505.3	21593	3.30%
Yamanishi_08	0.8–1.0	11.2	38.3	11044	0.44%
Yamanishi_08	0.6–0.8	18.7	96.8	18461	0.42%
Yamanishi_08	0.4–0.6	44.7	258.0	44175	0.59%
Yamanishi_08	0.2–0.4	76.5	531.8	75658	0.73%
Yamanishi_08	0.0–0.2	7.0	106.0	6923	1.23%

Extreme-support analyses based on the complete controlled model set retain only drugs with valid

Table 15: Support-resolved performance in Yamanishi_08 and ChEMBL37. Support-bin values are means over model–run evaluations; the final-column summary is identified in each cell. ChEMBL37 is reported as observed-positive recovery under positive–unlabelled semantics.

Dataset	$[0.8, 1.0]$	$[0.6, 0.8)$	$[0.4, 0.6)$	$[0.2, 0.4)$	$[0, 0.2)$	$S_{\max}$–AP Spearman
Yamanishi_08	0.484	0.381	0.264	0.094	0.023	0.486 (median)
ChEMBL37	0.542	0.452	0.280	0.076	0.034	0.627 (mean)

predictions from all eight models within the same run. The corresponding n_{ext} values can therefore be smaller than the descriptive counts in Table 14, particularly in the sparsest low-support cohorts.

Controlled model-set performance across support regimes. The support-resolved performance across evaluation regimes is summarized in Table 15.

To complement the aggregate support trends, we report the complete macro drug-level AP profiles of the controlled M1–M8 model set across evaluation regimes. Table 16 reports the MIFNDRA results, Table 17 reports Yamanishi_08, and Table 18 reports MDAD. Together, these profiles show how each model changes from Random Pair and overall Drug-Cold evaluation to progressively lower levels of training-set chemical support.

Table 16: Support-resolved performance of the controlled model set on MIFNDRA. Macro drug-level AP is reported across Random Pair and the available Drug-Cold chemical-support regimes, with values averaged across three independent runs.

ID	Model	Random	0.6–0.8	0.4–0.6	0.2–0.4	< 0.2
M1	Linear Logistic	0.055	0.021	0.064	0.043	0.053
M2	ExtraTrees	0.048	0.014	0.060	0.031	0.039
M3	Random Forest	0.056	0.039	0.064	0.066	0.054
M4	Two-Tower Dot Product	0.077	0.027	0.079	0.058	0.103
M5	Two-Tower MLP	0.090	0.031	0.068	0.057	0.066
M6	Bilinear Interaction	0.083	0.035	0.069	0.052	0.071
M7	Pair MLP	0.097	0.044	0.086	0.070	0.077
M8	kNN Similarity Transfer	0.035	0.012	0.062	0.022	0.041

Table 17: Support-resolved performance of the controlled model set on Yamanishi_08. Macro drug-level AP is reported across Random Pair and the available Drug-Cold chemical-support regimes, with values averaged across three independent runs.

ID	Model	Random	0.6–0.8	0.4–0.6	0.2–0.4	< 0.2
M1	Linear Logistic	0.083	0.080	0.080	0.063	0.024
M2	ExtraTrees	0.218	0.338	0.226	0.083	0.023
M3	Random Forest	0.237	0.335	0.247	0.079	0.023
M4	Two-Tower Dot Product	0.267	0.517	0.340	0.097	0.014
M5	Two-Tower MLP	0.304	0.556	0.401	0.125	0.017
M6	Bilinear Interaction	0.292	0.572	0.372	0.120	0.017
M7	Pair MLP	0.272	0.384	0.226	0.092	0.017
M8	kNN Similarity Transfer	0.156	0.269	0.222	0.090	0.051

Performance responses to reduced chemical support vary substantially across tasks. Yamanishi_08 shows a pronounced decline toward the extreme-support regime across the controlled model set, whereas MIFNDRA and MDAD exhibit more heterogeneous trajectories across models and support bins. Predictive degradation and changes in model ordering therefore do not necessarily coincide across datasets.

Table 18: Support-resolved performance of the controlled model set on MDAD. Macro drug-level AP is reported across Random Pair and the available Drug-Cold chemical-support regimes, with values averaged across three independent runs.

ID	Model	Random	0.6–0.8	0.4–0.6	0.2–0.4	< 0.2
M1	Linear Logistic	0.180	0.398	0.385	0.277	0.531
M2	ExtraTrees	0.202	0.496	0.444	0.317	0.486
M3	Random Forest	0.201	0.478	0.459	0.335	0.497
M4	Two-Tower Dot Product	0.204	0.490	0.468	0.336	0.559
M5	Two-Tower MLP	0.217	0.549	0.516	0.350	0.466
M6	Bilinear Interaction	0.210	0.473	0.459	0.328	0.501
M7	Pair MLP	0.177	0.366	0.402	0.314	0.497
M8	kNN Similarity Transfer	0.165	0.424	0.320	0.142	0.131

A.8 Covariate-Adjusted Support–Performance Associations

Chemical support may covary with other drug-level properties that are also associated with predictive performance. We therefore examine the relationship between per-drug $S_{\max}$ and AP before and after accounting for positive degree, molecular weight, and heavy-atom count. For each model–run combination, we first compute the Spearman correlation between $S_{\max}$ and AP and then estimate the corresponding rank-based partial association after adjustment for these covariates. Table 19 reports the median unadjusted and adjusted correlations across model–run combinations, together with the median p value for the adjusted association.

Table 19: Covariate-adjusted support–performance associations. Spearman correlations between per-drug $S_{\max}$ and AP are reported before and after adjustment for positive degree, molecular weight, and heavy-atom count. Values are medians across model–run combinations; p values correspond to the adjusted associations.

Dataset	Unadjusted ρ	Adjusted ρ	Median p
MDAD	0.134	0.161	0.0393
MIFNDRA	-0.103	0.123	0.458
Yamanishi_08	0.486	0.499	4.62e-11

On Yamanishi_08, the association remains strong after adjustment, with the median correlation changing only slightly from $\rho = 0.486$ to $\rho = 0.499$. In contrast, MIFNDRA and MDAD show substantially weaker associations. The relationship between chemical support and predictive performance is therefore dataset dependent rather than a consistent monotonic pattern across tasks.

A.9 Model-Order Transitions Across Chemical-Support Regimes

Model ordering changes substantially across evaluation regimes in several datasets. Table 20 reports the run-level transition from Random Pair to the extreme-support regime ($S_{\max} < 0.2$) for the controlled M1–M8 model set across the primary datasets, ChEMBL37, and the additional RNA–drug settings. The results reveal marked heterogeneity across tasks: Yamanishi_08, MDAD, and ChEMBL37 exhibit frequent changes in the highest-ranked model and weak rank agreement in multiple runs, whereas MIFNDRA retains substantially greater ordering consistency. MDS and MDR provide additional RNA–drug comparisons under the same evaluation protocol.

Within-Drug-Cold moderate-to-extreme transitions. To distinguish support-conditioned ordering changes from differences between split protocols, we compare model rankings entirely within

Table 20: Run-level model-order transitions from Random Pair to the extreme-support regime. The controlled M1–M8 models are ranked by macro drug-level AP within each independent run. The extreme-support regime is defined as $S_{\max} < 0.2$.

Dataset	Run	Spearman	Kendall	Top-1	Top-3	Random winner	Extreme winner
ChEMBL37	Run 1	0.333	0.214	1/1	0.333	M5	M5
ChEMBL37	Run 2	0.571	0.357	0/1	0.667	M5	M4
ChEMBL37	Run 3	-0.643	-0.500	0/1	0.000	M5	M1
MDAD	Run 1	-0.048	-0.071	0/1	0.000	M5	M1
MDAD	Run 2	0.167	0.143	0/1	0.333	M5	M4
MDAD	Run 3	0.429	0.429	0/1	0.667	M5	M3
MDR	Run 1	0.429	0.286	0/1	0.333	M4	M3
MDR	Run 2	0.714	0.500	0/1	0.667	M4	M5
MDR	Run 3	0.214	0.143	0/1	0.667	M7	M5
MDS	Run 1	0.714	0.500	0/1	0.667	M4	M7
MDS	Run 2	0.810	0.643	0/1	1.000	M7	M5
MDS	Run 3	0.857	0.714	0/1	0.667	M7	M6
MIFNDRA	Run 1	0.571	0.357	0/1	0.333	M5	M6
MIFNDRA	Run 2	0.905	0.786	0/1	1.000	M5	M7
MIFNDRA	Run 3	0.762	0.571	0/1	0.667	M7	M4
Yamanishi_08	Run 1	-0.619	-0.500	0/1	0.333	M5	M8
Yamanishi_08	Run 2	-0.952	-0.857	0/1	0.000	M5	M8
Yamanishi_08	Run 3	0.667	0.500	1/1	0.667	M5	M5

Drug-Cold. For each independent run, the controlled M1–M8 models are ranked by macro drug-level AP in the moderate-support regime, $S_{\max} \in [0.4, 0.6)$, and the extreme-support regime, $S_{\max} < 0.2$. Yamanishi_08, MDAD, and ChEMBL37 show low or negative rank agreement together with substantial pairwise reversals between these regimes. MIFNDRA has only 3/4/4 drugs in the moderate-support cohort across the three runs, resulting in substantially greater sampling uncertainty. Table 21 summarizes the aggregate transition statistics, and Table 22 reports the corresponding run-level results.

Table 21: Within-Drug-Cold model-order stability across chemical-support regimes. Each comparison ranks the controlled M1–M8 model set independently within each run and summarizes rank agreement, Top-k consistency, and meaningful pairwise reversals across three independent runs.

Dataset	Comparison	Spearman	Kendall's τ	Top-1	Top-3	Reversal rate 0.005	Reversal rate 0.01
Yamanishi_08	$[0.6, 0.8)$ vs. < 0.2	-0.262	-0.167	33.3%	0.222	47.6%	33.3%
	$[0.4, 0.6)$ vs. < 0.2	-0.373	-0.286	33.3%	0.222	48.8%	35.7%
	$[0.2, 0.4)$ vs. < 0.2	-0.111	-0.048	33.3%	0.333	36.9%	26.2%
MDAD	$[0.6, 0.8)$ vs. < 0.2	0.143	0.071	0.0%	0.333	39.3%	36.9%
	$[0.4, 0.6)$ vs. < 0.2	-0.008	-0.024	0.0%	0.111	44.0%	39.3%
	$[0.2, 0.4)$ vs. < 0.2	0.135	0.071	0.0%	0.444	34.5%	31.0%
ChEMBL37	$[0.6, 0.8)$ vs. < 0.2	0.040	0.071	33.3%	0.333	35.7%	28.6%
	$[0.4, 0.6)$ vs. < 0.2	0.087	0.024	33.3%	0.444	39.3%	32.1%
	$[0.2, 0.4)$ vs. < 0.2	0.310	0.238	33.3%	0.667	20.2%	15.5%
MIFNDRA	$[0.6, 0.8)$ vs. < 0.2	Insufficient cohort size (1/3/2 / 8/6/6)					
	$[0.4, 0.6)$ vs. < 0.2	Insufficient cohort size (3/4/4 / 8/6/6)					
	$[0.2, 0.4)$ vs. < 0.2	Insufficient cohort size (10/9/11 / 8/6/6)					

Figure 5 summarizes the same comparisons across progressively lower-support Drug-Cold regimes. The strongest loss of rank agreement occurs near the extreme-support tail in Yamanishi_08, MDAD, and ChEMBL37, accompanied by elevated rates of meaningful pairwise reversals.

Sensitivity to near ties. The observed transitions are not confined to model pairs separated by negligible performance differences. Increasing the minimum macro-AP difference required for a meaningful pairwise reversal from 0.005 to 0.01 reduces the number of eligible pairs, as expected, but substantial pairwise reversal rates remain in the extreme-support regime for Yamanishi_08, MDAD, and ChEMBL37. Table 23 reports the corresponding sensitivity results across support regimes.

Table 22: Run-level moderate-to-extreme ranking transitions within Drug-Cold. Model rankings compare the moderate-support regime, $S_{\max} \in [0.4, 0.6)$, with the extreme-support regime, $S_{\max} < 0.2$, using macro drug-level AP within each independent run.

Dataset	Run	$n_{0.4-0.6}$	$n_{<0.2}$	Spearman	Kendall	Moderate winner	Extreme winner	Reversal rate 0.005
Yamanishi_08	Run 1	43	6	-0.690	-0.500	M5	M8	60.7%
	Run 2	49	7	-0.881	-0.714	M6	M8	64.3%
	Run 3	40	2	0.452	0.357	M5	M5	21.4%
MDAD	Run 1	45	8	-0.238	-0.286	M5	M1	64.3%
	Run 2	36	9	-0.167	-0.071	M5	M4	50.0%
	Run 3	48	8	0.381	0.286	M2	M3	17.9%
ChEMBL37	Run 1	70	6	0.333	0.214	M5	M5	35.7%
	Run 2	68	8	0.476	0.286	M5	M4	28.6%
	Run 3	83	6	-0.548	-0.429	M5	M1	53.6%
MIFNDRA	Run 1	3	8	0.310	0.143	M7	M6	10.7%
	Run 2	4	6	0.714	0.571	M7	M7	7.1%
	Run 3	4	6	0.429	0.357	M7	M4	3.6%

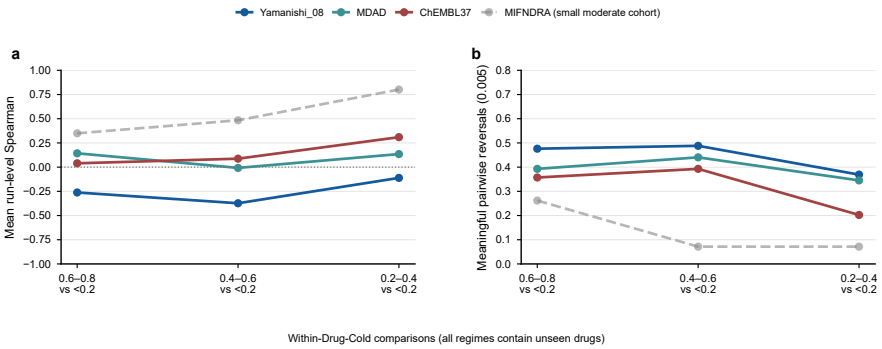


Figure 5: **Ranking transitions within Drug-Cold across chemical-support regimes.** (a) Mean Spearman rank agreement and (b) mean meaningful pairwise reversal rate between each moderate-support regime and $S_{\max} < 0.2$, averaged across three independent runs. Meaningful pairwise reversals require a macro-AP difference greater than 0.005 in both compared regimes. All comparisons use unseen drugs from the same Drug-Cold protocol. The dashed MIFNDRA trajectory denotes its small moderate-support cohorts.

Table 23: Sensitivity of ranking transitions to the near-tie threshold. Meaningful pairwise reversals are retained only when the model-pair macro-AP difference exceeds the specified threshold in both compared regimes.

Dataset	0.4–0.6 (0.005)	0.4–0.6 (0.01)	0.2–0.4 (0.005)	0.2–0.4 (0.01)	< 0.2 (0.005)	< 0.2 (0.01)
MIFNDRA	9.5%	2.4%	11.9%	7.1%	9.5%	6.0%
Yamanishi_08	11.9%	9.5%	16.7%	8.3%	48.8%	35.7%
MDAD	11.9%	3.6%	10.7%	4.8%	34.5%	23.8%
ChEMBL37	1.2%	1.2%	6.0%	2.4%	34.5%	25.0%

Rank trajectories across support regimes. Figure 6 traces the mean rank of each controlled model across progressively lower levels of training-set chemical support. The trajectories make clear that rank changes are not restricted to a single transition point: several models exchange positions across intermediate and low-support regimes, with the magnitude and direction of these changes varying by dataset.

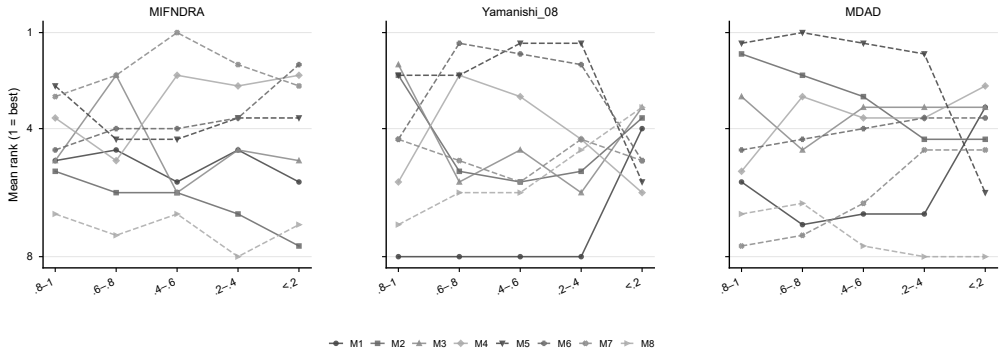


Figure 6: **Controlled-model rank trajectories across chemical-support regimes.** Mean ranks of M1–M8 are shown across the pre-specified Drug-Cold support intervals; lower values indicate better performance. Marker and line-style encodings identify individual models.

A.10 RNA–Drug Robustness Across Additional Datasets

We extend the support-conditioned ranking analysis to two additional RNA–drug datasets, MDS and MDR, using the same controlled M1–M8 model set, partitioning protocol, chemical-support definition, and evaluation metrics. This provides an additional assessment of model-order stability within the same interaction-task family while varying the underlying dataset.

The two datasets exhibit distinct stability patterns. MDR shows substantial run-to-run variation in Random Pair-to-extreme-support rank agreement, together with non-negligible meaningful pairwise reversals. MDS is considerably more stable, with consistently positive rank correlations and lower reversal rates across the three independent runs. Table 24 reports the corresponding run-level Spearman and Kendall correlations together with meaningful pairwise reversal rates under both the 0.005 and 0.01 macro-AP thresholds.

Table 24: Run-level ranking stability on the additional RNA–drug datasets. Spearman and Kendall rank correlations compare Random Pair with the extreme-support regime ($S_{\max} < 0.2$), while meaningful pairwise reversal rates are reported under both the 0.005 and 0.01 macro-AP thresholds.

Dataset	Run	Spearman	Kendall	Reversal rate 0.005	Reversal rate 0.01
MDR	Run 1	0.429	0.286	35.7%	32.1%
MDR	Run 2	0.714	0.500	17.9%	14.3%
MDR	Run 3	0.214	0.143	25.0%	17.9%
MDS	Run 1	0.714	0.500	0.0%	0.0%
MDS	Run 2	0.810	0.643	14.3%	10.7%
MDS	Run 3	0.857	0.714	3.6%	0.0%

A.11 ChEMBL37 Protocol

ChEMBL37 is evaluated as a direct-human drug–protein dataset under positive–unlabelled semantics. Experimentally observed direct-human interactions are treated as positives, whereas unobserved drug–protein candidates remain unlabelled. For ranking-based evaluation, unobserved candidates are encoded as 0 without being interpreted as confirmed negatives.

The ChEMBL37 experiments use the same chemical representation, pre-specified $S_{\max}$ intervals, controlled model families, and three independent runs as the primary evaluation. Performance is reported as macro observed-positive AP over candidate drug–protein pairs. Support-resolved profiles describe observed-positive recovery and model-order variation across levels of training-set chemical support.

A.12 Uncertainty in Extreme-Support Ranking Stability

The extreme-support regime is defined by the pre-specified threshold $S_{\max} < 0.2$ and contains relatively few test drugs in several datasets. Within each dataset–run cohort, we perform 1,000 whole-drug bootstrap resamples with replacement. For each replicate, macro drug-level AP is recomputed from the resampled per-drug predictions for all eight controlled models, and the resulting model ordering is compared with the fixed Random Pair ordering. Rank-agreement and pairwise-reversal statistics are averaged across the three independent runs. Resampling is performed at the drug level, consistent with the macro drug-level evaluation unit.

Bootstrap estimates show pronounced Random Pair-to-extreme ordering changes in Yamanishi_08, MDAD, and ChEMBL37, although the corresponding intervals are broad because the extreme-support cohorts contain only two to nine drugs per run. MIFNDRA shows comparatively stable ordering with positive rank agreement. Table 25 reports the observed statistics and bootstrap intervals, and Figure 7 shows the corresponding uncertainty in rank correlation and meaningful pairwise reversals.

Table 25: Bootstrap uncertainty in extreme-support model ordering. Observed rank correlations and meaningful pairwise reversal rates are compared with whole-drug bootstrap estimates from 1,000 resampling replicates. The 95% confidence interval (CI) columns summarize uncertainty across the extreme-support cohorts, and the final columns report the stability of Random Pair and observed extreme-support Top-1 identities.

Dataset	Mean n_{extreme}	Observed ρ	Bootstrap ρ [95% CI]	Observed reversals	Bootstrap reversals [95% CI]	Pr(Random)	Pr(observed)
Yamanishi_08	5.0	−0.302	−0.230 [−0.437, 0.207]	48.8%	40.5% [19.0, 53.6]%	24.5%	75.7%
MDAD	8.3	0.183	0.230 [−0.024, 0.502]	34.5%	32.1% [20.2, 44.0]%	0.0%	64.4%
ChEMBL37	6.7	0.087	0.016 [−0.453, 0.373]	34.5%	33.3% [16.7, 40.5]%	18.2%	49.8%
MIFNDRA	6.7	0.746	0.694 [0.500, 0.833]	9.5%	11.9% [3.6, 20.2]%	0.0%	85.5%

Leave-one-drug-out analysis yields the same qualitative pattern. In primary runs where the Random Pair and extreme-support winners differ, removing any single extreme-support drug does not restore the Random Pair winner. In one Yamanishi_08 run and one ChEMBL37 run, the two winners are already identical before removal. The Top-1 transitions are therefore not driven by a single extreme-support compound, although winner identity remains variable within these small cohorts.

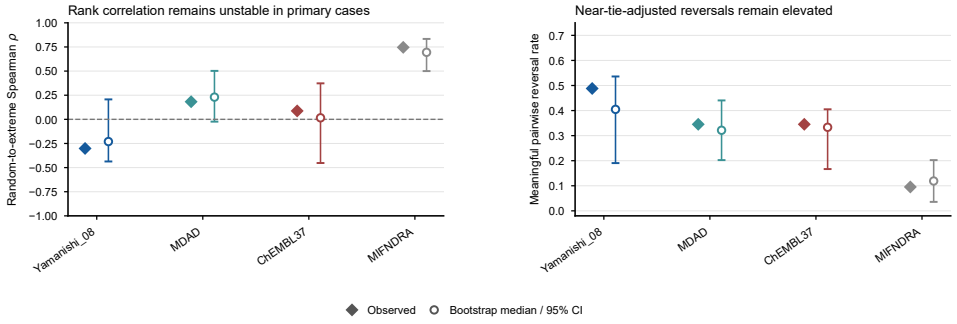


Figure 7: **Bootstrap uncertainty in extreme-support ranking stability.** Diamonds denote observed three-run means; open circles and vertical intervals denote bootstrap medians and 95% intervals from 1,000 whole-drug resampling replicates. Drugs are resampled within each run before statistics are averaged across runs, with the Random Pair ordering held fixed. Meaningful pairwise reversals require a macro-AP difference greater than 0.005 in both compared regimes. MIFNDRA is shown in grey because its ordering remains comparatively stable.

Equal-size moderate-support comparison. Within each dataset–run, 1,000 subsets are sampled without replacement from the moderate-support Drug-Cold cohort, $S_{\max} \in [0.4, 0.6)$, with each subset matched in size to the observed $S_{\max} < 0.2$ cohort. The comparison therefore uses the same model set, run, evaluation protocol, and number of drugs while differing in training-set chemical support.

Extreme-support rankings are less stable than the corresponding equal-size moderate-support distributions in Yamanishi_08, MDAD, and ChEMBL37. One-sided Monte Carlo comparisons yield lower rank correlations and higher meaningful pairwise reversal rates in the extreme-support cohorts. After Holm correction across the three evaluable datasets, adjusted p values are at most 0.011 for rank correlation and 0.013 for reversal rate. Table 26 reports the observed extreme-support statistics together with the equal-size moderate-support reference distributions. MIFNDRA is not evaluable because its moderate-support pool is smaller than its extreme-support cohort.

Table 26: Equal-size moderate-support comparison for extreme-support ranking instability. For each dataset–run, 1,000 subsets are sampled from the $S_{\max} \in [0.4, 0.6)$ Drug-Cold cohort with the same number of drugs as the observed $S_{\max} < 0.2$ cohort. The table compares extreme-support rank correlations and meaningful pairwise reversal rates with the corresponding equal-size moderate-support reference distributions; reported p values include Holm-adjusted results across evaluable datasets.

Dataset	n_{ext}	Extreme ρ	Moderate-support reference ρ [95%]	p_ρ	$p_{\text{H},\rho}$	Extreme reversals	Moderate-support reference reversals [95%]	p_{rev}	$p_{\text{H},rev}$
Yamanishi_08	5.0	-0.302	0.604 [0.222, 0.833]	0.001	0.003	48.8%	23.1% [11.9, 38.1]%	0.001	0.003
MDAD	8.3	0.183	0.540 [0.254, 0.826]	0.011	0.011	34.5%	21.3% [8.3, 33.3]%	0.013	0.013
ChEMBL37	6.7	0.087	0.732 [0.389, 0.921]	0.001	0.003	34.5%	14.5% [4.8, 28.6]%	0.004	0.008
MIFNDRA	6.7	Not evaluable: moderate pool smaller than extreme cohort in Runs 1–3 (3/4/4 vs. 8/6/6 drugs)							

A.13 Oracle Gap and Label-Disjoint Model Selection

The ex-post oracle gap is defined as the extreme-support macro drug-level AP of the best-performing model minus that of the Random Pair winner. Positive values indicate a change in the top-performing model under $S_{\max} < 0.2$. Because model identification and evaluation use the same extreme-support labels, the gap is retrospective. Table 27 reports the observed gaps and whole-drug bootstrap intervals.

Table 27: Ex-post oracle gap in the extreme-support regime. The gap is the macro drug-level AP of the best-performing extreme-support model minus that of the Random Pair winner. Values include whole-drug bootstrap medians and 95% CIs.

Dataset	Observed	Bootstrap median	95% CI
Yamanishi_08	0.038	0.039	[0.009, 0.093]
MDAD	0.130	0.148	[0.084, 0.240]
ChEMBL37	0.032	0.035	[0.005, 0.097]
MIFNDRA	0.048	0.046	[0.012, 0.103]

Label-disjoint selection evaluates model choice on non-overlapping subsets of extreme-support drugs. Each extreme-support cohort is partitioned into balanced selection and evaluation subsets; the model with the highest macro drug-level AP on the selection subset is then evaluated on the disjoint subset against the Random Pair winner. All balanced partitions are enumerated, with both complementary orientations retained when the cohort size is odd. Table 28 reports the resulting held-out macro-AP differences across all partitions, providing a label-disjoint measure of how consistently support-aware model selection transfers across extreme-support compounds.

Table 28: Label-disjoint model selection in the extreme-support regime. Balanced, non-overlapping drug subsets are used for model selection and evaluation. Δ denotes the held-out macro drug-level AP of support-aware selection minus that of the Random Pair winner.

Dataset	n_{ext}	Mean Δ	5th–95th Δ	$\Pr(\Delta > 0)$
ChEMBL37	6.7	-0.016	[-0.046, 0.005]	19.8%
MDAD	8.3	0.059	[-0.062, 0.160]	74.3%
MIFNDRA	6.7	0.029	[-0.005, 0.071]	71.9%
Yamanishi_08	5.0	0.010	[-0.006, 0.043]	56.4%

A.14 Robustness Across Representation, Metric, and Cohort Definitions

Representation sensitivity. Support-conditioned ranking patterns are broadly preserved under alternative molecular representations, although the composition of the lowest-support tail is representation dependent. With radius-3 Morgan fingerprints, Yamanishi_08 retains negative Random Pair-to-extreme rank agreement ($\rho = -0.246$), while the meaningful pairwise reversal rates in MDAD and ChEMBL37 remain 34.5% at the 0.005 macro-AP threshold. MACCS-166 substantially changes the composition of the low-support tail and yields weaker ranking transitions in MDAD and ChEMBL37, with rank correlations of 0.667 and 0.556 and reversal rates of 11.9% in both datasets. Yamanishi_08 remains less stable, with $\rho = 0.333$ and 31.0% meaningful pairwise reversals. Table 29 reports the corresponding agreement in per-drug $S_{\max}$ values and bottom-decile overlap across representations.

Table 29: Agreement across chemical-support representations. The table compares per-drug $S_{\max}$ values and bottom-decile overlap between radius-2 Morgan, radius-3 Morgan, and MACCS-166 representations. Values are averaged across three independent runs.

	Morgan r2 vs Morgan r3		Morgan r2 vs MACCS-166	
Dataset	$S_{\max}$ ρ	Bottom-10% overlap	$S_{\max}$ ρ	Bottom-10% overlap
Yamanishi_08	0.987	83.3%	0.725	27.1%
MDAD	0.978	82.4%	0.717	49.0%
ChEMBL37	0.984	71.9%	0.601	25.0%
MIFNDRA	0.979	88.9%	0.685	44.4%

Metric sensitivity. Model ordering is also examined using mean Precision@20 and Hit@20 in addition to the primary macro drug-level AP metric. The same per-drug predictions are re-ranked under each metric, and Random Pair-to-extreme rank agreement and Top-1 consistency are compared across datasets. Yamanishi_08 and ChEMBL37 retain near-zero rank agreement under both alternative metrics, while MDAD shows higher correlations but zero Top-1 agreement. MIFNDRA remains comparatively stable. Because the 0.005 near-tie threshold is defined on the macro-AP scale, it is not applied to Precision@20 or Hit@20. The resulting metric-specific ranking statistics are reported in Table 30.

Table 30: Model-order stability across evaluation metrics. Random Pair-to-extreme rank correlation and Top-1 agreement are reported under macro drug-level AP, Precision@20, and Hit@20.

	Macro drug AP		Precision@20		Hit@20	
Dataset	ρ	Top-1	ρ	Top-1	ρ	Top-1
Yamanishi_08	-0.302	0.333	0.043	0.333	0.018	0.000
MDAD	0.183	0.000	0.431	0.000	0.366	0.000
ChEMBL37	0.087	0.333	0.010	0.333	-0.017	0.000
MIFNDRA	0.746	0.000	0.554	0.000	0.531	0.000

Matched-cohort sensitivity. For MDAD, each extreme-support drug ($S_{\max} < 0.2$) is matched within run to one moderate-support drug ($S_{\max} \in [0.4, 0.6)$) using positive count, positive prevalence, molecular weight, and heavy-atom count. The mean absolute standardized differences af-

ter matching are 0.050, 0.050, 0.087, and 0.103, respectively. Despite this covariate balance, the matched model orderings retain low agreement (mean $\rho = 0.087$), a 42.9% meaningful pairwise reversal rate, and zero Top-1 agreement across the three independent runs.

Relative-tail sensitivity. Representation-specific bottom-5% and bottom-10% cohorts provide complementary definitions of low chemical support without relying on a fixed absolute threshold. These cohorts are constructed separately using radius-2 Morgan, radius-3 Morgan, and MACCS-166 while retaining the same Drug-Cold memberships and per-drug predictions. Table 31 reports the resulting cohort sizes, tail boundaries, rank correlations, and pairwise reversal rates.

Table 31: Ranking stability in representation-specific low-support tails. Bottom-5% and bottom-10% cohorts are defined separately under radius-2 Morgan, radius-3 Morgan, and MACCS-166 using the same Drug-Cold test drugs. Values are averaged across three independent runs.

Dataset	Representation	Bottom fraction	Drugs	Tail upper $S_{\max}$	ρ	Reversals
ChEMBL37	MACCS-166	5%	16.0	0.622	0.944	3.6%
ChEMBL37	MACCS-166	10%	32.0	0.650	0.802	4.8%
ChEMBL37	Morgan r2	5%	16.0	0.221	0.294	26.2%
ChEMBL37	Morgan r2	10%	32.0	0.242	0.468	15.5%
ChEMBL37	Morgan r3	5%	16.0	0.174	0.444	17.9%
ChEMBL37	Morgan r3	10%	32.0	0.185	0.444	20.2%
MDAD	MACCS-166	5%	9.0	0.563	0.563	17.9%
MDAD	MACCS-166	10%	17.0	0.643	0.540	19.0%
MDAD	Morgan r2	5%	9.0	0.199	0.135	35.7%
MDAD	Morgan r2	10%	17.0	0.249	0.452	23.8%
MDAD	Morgan r3	5%	9.0	0.155	0.230	33.3%
MDAD	Morgan r3	10%	17.0	0.188	0.381	27.4%
MIFNDRA	MACCS-166	5%	2.0	0.416	0.103	42.9%
MIFNDRA	MACCS-166	10%	3.0	0.495	0.190	36.9%
MIFNDRA	Morgan r2	5%	2.0	0.159	0.587	13.1%
MIFNDRA	Morgan r2	10%	3.0	0.171	0.563	11.9%
MIFNDRA	Morgan r3	5%	2.0	0.128	0.587	13.1%
MIFNDRA	Morgan r3	10%	3.0	0.129	0.571	11.9%
Yamanishi_08	MACCS-166	5%	8.0	0.565	0.294	28.6%
Yamanishi_08	MACCS-166	10%	16.0	0.599	0.754	15.5%
Yamanishi_08	Morgan r2	5%	8.0	0.213	-0.397	54.8%
Yamanishi_08	Morgan r2	10%	16.0	0.236	-0.302	50.0%
Yamanishi_08	Morgan r3	5%	8.0	0.167	-0.429	56.0%
Yamanishi_08	Morgan r3	10%	16.0	0.183	-0.349	50.0%

A.15 Summary of Empirical Evidence

Across the analyses, scaffold novelty and continuous chemical support emerge as distinct evaluation properties, and reduced chemical support does not induce a uniform performance response across datasets. Model ordering changes substantially in several low-support settings, although the magnitude of these transitions varies with the dataset, candidate model set, molecular representation, evaluation metric, and cohort definition. These conclusions are supported by convergent evidence from within-Drug-Cold comparisons, equal-size controls, whole-drug resampling, label-disjoint selection, and representation- and metric-level sensitivity analyses. Table 32 summarizes the main empirical findings together with their supporting analyses and scope conditions.

Table 32: Summary of empirical findings and supporting analyses. Each row links a main finding to the corresponding primary analysis, robustness checks, and scope conditions.

Finding	Primary analysis	Sensitivity analysis	Scope
Scaffold novelty does not define calibrated chemical support	Scaffold-Cold versus Drug-Cold $S_{\max}$ distributions	$S_{\max}$-matched macro-AP comparison	MDAD has limited matching overlap; support is representation dependent
Ranking can change in low-support regimes	Random Pair-to-$S_{\max} < 0.2$ ordering	Within-Drug-Cold comparison, equal-size moderate-support reference, whole-drug bootstrap, and leave-one-drug-out analysis	MIFNDRA has limited moderate-bin sample support for the within-Drug-Cold comparison; additional DTI architectures show weaker but non-zero support-conditioned ordering changes whose magnitude varies by dataset
Model-selection consequences are task dependent	Ex-post opportunity gap	Label-disjoint selection simulation	ChEMBL37 is positive–unlabelled; the selection analysis is retrospective and finite-cohort
Support definition is representation dependent	Radius-2 Morgan support	Radius-3 Morgan, MACCS-166, and bottom-tail sensitivity	Support thresholds are representation specific