RETROTAC: LEARNING ROUTE-DERIVED SYNTHETIC ACCESSIBILITY FOR PROTACS

Anonymous authors
Paper under double-blind review

ABSTRACT

Proteolysis-targeting chimeras (PROTACs) enable targeted protein degradation, expanding therapeutic opportunities beyond conventional inhibition. However, assembling their large, multifunctional structures creates synthetic challenges that hinder their discovery. Traditional synthetic-accessibility scores can overestimate this difficulty by penalizing PROTACs' molecular complexity while overlooking modular assembly from accessible building blocks. In contrast, retrosynthesis planning tools provide explicit routes but are computationally expensive, particularly for large molecules. Here, we introduce a human-interpretable, PROTAC-oriented route-derived score that rewards short synthetic depth and balanced joining of molecular fragments, and RETROTAC, a machine-learning surrogate that predicts this score directly from molecular structures. RETROTAC combines complementary models in an ensemble, achieving a root mean squared error of 0.13 and an R^2 of 0.65 on held-out molecules. We further evaluate its utility on PROTACs enumerated by recombining warheads, linkers, and E3-ligase ligands, followed by retrosynthetic planning and route analysis. Candidates selected for high predicted scores achieve a mean route-derived synthesizability of 0.73, compared with 0.31 for low-score candidates; 98.75% versus 51.00% are fully resolved under the tested search settings. These results support RETROTAC as a fast prioritization tool for virtual PROTAC libraries, connecting molecular screening with interpretable route characteristics while reserving explicit planning for selected candidates.

1 INTRODUCTION

Proteolysis-targeting chimeras (PROTACs) are heterobifunctional molecules that recruit a protein of interest to an E3 ubiquitin ligase, inducing target degradation through the ubiquitin–proteasome system (Hu & Crews, 2022; Gharbi & Mercado, 2024). A PROTAC contains a target-binding warhead and an E3-ligase ligand connected by a linker. This modular architecture offers a route to modulating proteins that are difficult to drug with occupancy-driven inhibitors, but it also produces molecules that are larger and more flexible than conventional small-molecule drugs (Weerakoon et al., 2022; Price et al., 2024). Synthesis is therefore a practical bottleneck, especially as generative models can enumerate virtual degraders much faster than routes can be assessed or executed (Nori et al., 2022; Li et al., 2023; Guo et al., 2023; Xu et al., 2026).

Fast synthetic-accessibility scores are attractive filters, but they prioritize different properties of synthesis routes. Fragment-frequency heuristics assess whether a structure resembles common chemistry; reaction-corpus models infer relative complexity or feasibility; and retrosynthesis-derived classifiers estimate whether one planner succeeds under fixed settings (Ertl & Schuffenhauer, 2009; Voršilák et al., 2020; Chen & Jung, 2024; Coley et al., 2018; Neeser et al., 2024; Thakkar et al., 2021). Most were developed primarily in small-molecule-drug-like chemical space, so their size penalties, fragment vocabularies and reaction corpora may mischaracterize large molecules assembled by joining a small number of modular fragments, as in a PROTAC. Two efforts target PROTACs more directly: DeepPSA is trained specifically on PROTAC synthetic-accessibility data (Zhang et al., 2025), and FSscore has been fine-tuned on a small set of expert-labeled PROTAC pairs (Neeser et al., 2024).

Retrosynthetic planning tools provide richer evidence—including step count, convergence, proposed transformations and terminal building blocks—but cost substantially more than a single property

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.

prediction and cannot be easily integrated in generative pipeline or virtual screening campaigns. Their cost is even greater when processing large, complex molecules like PROTACs. Moreover, found retrosynthesis routes are conditional on the planner, reaction templates, stock, search budget and stochastic seed (Chen et al., 2020; Thakkar et al., 2020; Maziarz et al., 2025). We therefore reframe route assessment as single-molecule property prediction. We estimate synthesizability directly from a target PROTAC and without calling the planner at inference, using a transparent heuristic designed to summarize a route from a fixed external pipeline.

Our contributions to address the above limitations are threefold:

- we define a human-interpretable and PROTAC-oriented route-derived score that rewards short synthesis depth and balanced joining of reactants;
- we develop and evaluate a fast and accurate proxy model for retrosynthesis called RETROTAC;
- and we openly release nearly 20,000 PROTAC synthesis routes generated by ShallowTree, as well as the RETROTAC ensemble model.

2 RELATED WORK

Two prior scores address PROTACs directly. FSscore fine-tunes a general feasibility ranker on pairwise expert preferences over PROTAC-DB (Neeser et al., 2024; Weng et al., 2023); tuning shifts its scores toward the easy end and narrows the gap between a PROTAC and its most complex fragment, consistent with joining the three modules not being the hard step, but it draws on only tens of expert-labeled pairs. DeepPSA (Zhang et al., 2025), the only model so far built for PROTAC synthetic accessibility, is a graph classifier that decides easy versus hard over a proprietary set of 3,644 PROTACs. Its authors release trained weights, so we can run it on our own molecules (Appendix E), though its accuracy does not compare directly with a route-derived score.

Route scores draw on information molecule-only models lack, such as graph-edit distance with a length penalty (Gao et al., 2025), template n-gram similarity to known routes (Li et al., 2024), step penalties with forward-model confidence (Kreutter & Reymond, 2023), and recursive precursor cost with expert grades (Sun et al., 2021). Ours is the most lightweight of these route scores, and, to our knowledge, the first defined on PROTAC syntheses. It reads only route topology and heavy-atom counts, needs no second reaction model, and—unlike a length-only penalty—separates routes that decorate one large intermediate from routes that join comparably sized modules. While it doesn't capture feasibility, selectivity or yield, it does measure route depth and modular balance.

3 METHODS

We frame the learning objective as a regression task. Each sample contains a target PROTAC SMILES x_i and a route r_i supplied by an external retrosynthesis tool. The route determines the training label, which we refer to throughout as the route-derived score, but the predictor receives only the target molecule and learns $f_\theta(x_i) \approx y(r_i)$ (Figure 1a). We use this to create a surrogate for route planning: explicit routes are required to create training data, but not to score a new molecule.

3.1 DATA CURATION

We collected and standardized 21,189 PROTAC SMILES from PROTAC-DB (Ge et al., 2025), PROTACpedia (London & Prilusky, 2024), and TPDdb (Qin et al., 2026). We generated synthesis routes for these using ShallowTree, resulting in 18,076 PROTACs with route-derived labels. We used 15,844 molecules for model development and set aside 2,232 ($\sim$10%) for a final held-out evaluation. The held-out set was isolated before model fitting with adaptive Butina clustering (Butina, 1999) on 2,048 Morgan fingerprints, radius 8, at a Tanimoto-distance (Lipkus, 1999) cutoff of 0.85 (Appendix A.1). Figure 1c illustrates the RETROTAC training and evaluation pipeline. The development set (i.e., train and validation sets) requires a second safeguard against data leakage before cross-validation. PROTACs are assembled from recurring warheads, linkers and E3 ligands; we therefore decomposed every molecule with PROTAC-Splitter (Ribes et al., 2026) and grouped the development folds by Bemis–Murcko scaffolds (Bemis & Murcko, 1996) of the warhead (Appendix B). For constructing

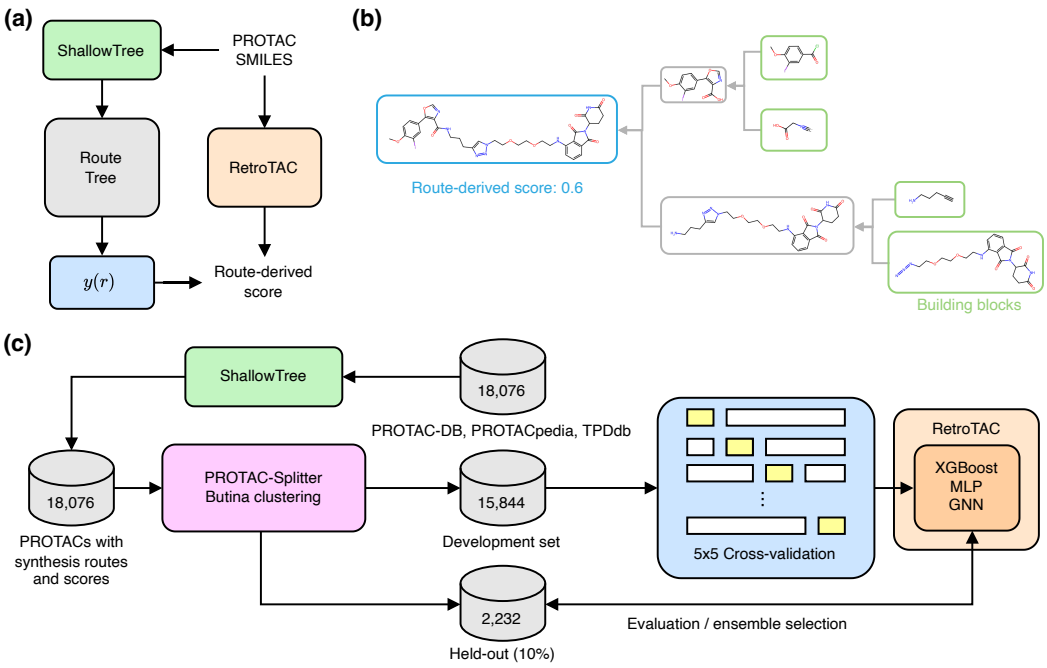


Figure 1: (a) RETROTAC is trained to proxy the route-derived score computed from synthesis routes found by ShallowTree. (b) Example synthesis route found by ShallowTree. (c) Training and evaluation pipeline. After obtaining the synthesis routes, we isolate a held-out set via Butina clustering. The remaining development set is used to train and compare three surrogate models in a repeated cross-validation setting.

the Caruana ensemble (Section 3.5) and comparing the other strategies, we isolated a hill-climb set of 463 (20%) randomly sampled data from the held-out set, and evaluated on the remaining 1,854 data points.

3.2 RETROSYNTHESIS ROUTE PREDICTION

In order to produce the synthetic routes for each of the PROTACs datapoints we used a proprietary retrosynthesis planning tool ShallowTree. ShallowTree is a synthesis planning tool that can generate synthetic routes for a query molecule at a user defined depth. It is inspired by the tools for single-molecule retrosynthesis guided by neural network expansion policy such as AiZynthFinder (Genheden et al., 2020). Depending on user's preferences, the search algorithm can balance depth first over breadth first search. ShallowTree expansion and filter policies as well as the building block collections are compatible with the ones used by AiZynthFinder. However, ShallowTree is designed for the purpose of processing multiple target molecules in parallel while also keeping a minimal memory and optimal CPU usage. This makes it particularly suitable for processing rather challenging targets such as PROTACs even when the dataset consists of tens of thousands of molecules. All searches used the eMolecules building-block collection as purchasable stock (eMolecules, 2026). We processed the entire dataset data set and successfully resolved about 92.4% or 17,802 out of 19,274 PROTACs molecules by using 64 CPUs with a timeout limit of 30 min per molecule. The maximum permitted depth was 12. The remaining 1,472 unresolved molecules exceeded the allocated timeout or exhausted the available trees at depth 12. Generating the synthetic trees for the entire dataset on 64 CPUs took less than 50h. An example of a synthetic tree generated by ShallowTree can be found on Figure 1b, and additional routes in Appendix F.

3.3 ROUTE-DERIVED SYNTHESIZABILITY SCORE DEFINITION

We designed the route-derived score around two key concerns in PROTAC synthesis: *how many sequential reactions are required before all branches reach purchasable building blocks*, and, *when*

the route joins fragments, does it combine similarly sized modules or iteratively decorate one large intermediate with small groups? The first property is represented by the longest linear sequence in the route, also known as *route depth*, denoted here as $D(r)$, which decreases linearly to zero at a fixed saturation depth $D_{\max}$. The second is represented by a balance term $B(r)$: at each bimolecular reaction step, we compare the heavy-atom counts of the joined fragments using the ratio of the smaller to the larger fragment, and then average over the route. The final score is

$$y(r) = \frac{1}{2}\left[\max\left(0, 1 - \frac{D(r)-1}{D_{\max}-1}\right) + B(r)\right], \qquad D_{\max} = 11. \tag{1}$$

Both the depth component and $B(r)$ lie in $[0, 1]$, where larger values describe shorter, more balanced routes. This two-term form emerged from a simpler design choice rather than a fitted optimization. A detailed analysis of the scoring design is reported in Appendix C.

Two boundary cases do not have a route to score. Targets already present in the purchasable stock receive $y = 1$, whereas searches that fail to reach stock within the planner budget receive $y = 0$. These endpoints are not symmetric: one records known availability, while the other may reflect either genuine difficulty or a limitation (e.g., timeout, or max depth reached) of the search.

3.4 Selected models and statistical comparison

We explored three model architectures for predicting the route-derived score (y): XGBoost (XGB) (Chen & Guestrin, 2016), multilayer perceptron (MLP), and Chemprop (GNN) (Graff et al., 2026). Both XGB and MLP receive 1,024-bit, radius-8 Morgan fingerprints with chirality enabled, alongside 2D RDKit descriptors. Fingerprints are eventually reduced with truncated SVD via hyperparameters optimization, while descriptors are median-imputed, constant columns are removed, and the remaining values are standardized; the GNN instead receives the molecular graph parsed directly from SMILES. See Appendix C.1 for details.

To compare the three model backbones, we follow the repeated-cross-validation (CV) framework of Ash et al. (2025), giving 25 paired evaluations per model. For every outer fold, hyperparameters and early stopping are chosen on a separate grouped by warhead scaffolds 80/20 split of the available training molecules; the outer fold remains untouched until final scoring. We then refit each model on the full development set and evaluate it once on the held-out set.

Statistical comparisons of the 25 performance measurements from CV are performed with AutoRank (Herbold, 2020) at a significance level of 0.05. To select the testing procedure, we assess between-model differences in fold-score variance using Levene's test and the ratio of the largest to the smallest variance. We flag substantial heteroscedasticity only when Levene's $p < 0.05$ and this ratio exceeds 9 (Blanca et al., 2018). Otherwise, we explicitly select AutoRank's parametric procedure: repeated-measures ANOVA followed by Tukey HSD for pairwise comparisons with family-wise error control. This branch assumes approximate normality regardless of the normality diagnostic results. When the variance criterion is exceeded, we retain AutoRank's automatic test selection based on its normality and variance-homogeneity checks.

We evaluate regression performance primarily using R^2 and RMSE, complemented by MAE for absolute prediction error and Spearman's ρ for ranking performance. For a secondary classification analysis, we treat route-derived scores of ≥ 0.7 as positive and apply the same threshold to model predictions. These binary labels are used to compute MCC, F1, precision, recall, specificity, balanced accuracy, accuracy, Cohen's κ. ROC-AUC and PR-AUC instead use continuous predictions against the binary reference labels.

3.5 Ensemble selection

After identifying the best performing models, we compared three strategies for selecting cross-validated models for an ensemble: uniform weighting of all 5×5 CV models and architectures (XGB, MLP, and GNN), for a total of 75 models; the best backend architecture among XGB, MLP, or GNN, with 25 models; and a Caruana et al. (2004) greedy selection over all the 75 models. The ensembles return mean predictions and uncertainty, measured as standard deviation of individual models' outputs. We refer to the best ensemble model as RetroTAC throughout this work.

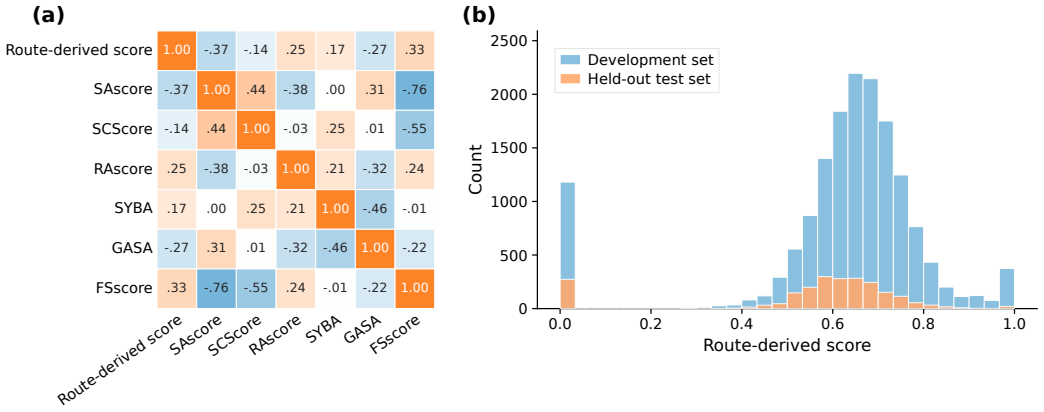


Figure 2: (a) Spearman correlation matrix between our proposed route-derived score and common synthesizability scores developed for small molecules. (b) Development versus held-out set distribution of the route-derived synthesizability score.

3.6 Prospective validation

We scored the unseen, enumerated PROTACs from Ribes et al. (2026) using the RetroTAC surrogate, recording each molecule's predicted synthesizability (the Caruana-weighted ensemble mean) and its uncertainty (the unweighted standard deviation across ensemble members). We predicted 1,303,154 PROTACs in total, after removing 71 samples overlapping with the development set. We divided the scored molecules by uncertainty at the median into a *confident* pool and an *uncertain* pool. Within each pool, we isolated the 400 lowest- and 400 highest-scoring molecules by predicted synthesizability, yielding four subsets: confident-low and confident-high, and uncertain-low and uncertain-high. We ran ShallowTree on each subset to obtain a potential synthesis route per molecule, scored these routes as in Section 3.3, and compared to the original RetroTAC predictions.

4 Results

4.1 Score distributions and correlations

In Figure 2a we compare the correlation between our proposed route-derived score and existing synthesizability scores calculated on all the collected PROTACs (SA score, $\rho = -0.37$; SCScore, $\rho = -0.14$; RAscore, $\rho = 0.25$; SYBA, $\rho = 0.17$; GASA, $\rho = -0.27$; and FSscore, $\rho = 0.33$) illustrating weak correlation, and even anti-correlation, with existing metrics, suggesting current metrics which have been mostly designed for small molecule drugs do not transfer well to the PROTAC space. We next examine the route-derived synthesizability score developed herein across the development ($n = 15,844$) and held-out ($n = 2,232$) sets. The held-out distribution was slightly shifted towards lower values, with a larger concentration at zero (Figure 2b); a two-sample Kolmogorov–Smirnov test confirmed the two distributions differ significantly ($D = 0.1807$, $p = 2.513 \times 10^{-56}$), likely due to the peaks at 0.

4.2 Cross-validation and held-out performance

The GNN achieved the highest mean R^2 (0.509 ± 0.056) and lowest RMSE (0.142 ± 0.010), followed by MLP (0.290 ± 0.114 and 0.171 ± 0.016, respectively) and XGB (0.237 ± 0.121 and 0.177 ± 0.018). Results are summarized in Table 1. The GNN also achieved the lowest mean MAE (0.094) and highest mean Spearman correlation (0.597), with lower variability across folds in R^2 and RMSE than either baseline.

To compare the R^2 from different model backbones, Levene's test ($p = 0.0877$) and the maximum-to-minimum fold-variance ratio (4.68) met our criterion for selecting the parametric procedure, although the Shapiro-Wilk test rejected normality for MLP ($p = 0.001$). Under this procedure, repeated-measures ANOVA detected differences in mean R^2 ($p < 0.001$), and Tukey HSD identified

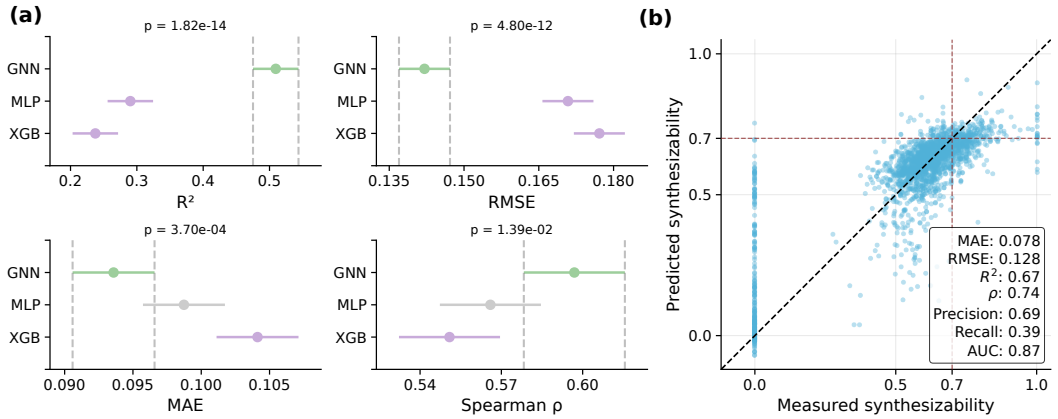


Figure 3: (a) Simultaneous confidence interval plots comparing the three models on regression metrics. (b) Parity plots of the Caruana-based ensemble.

Table 1: Regression, ranking, and thresholded-classification performance. CV values are mean ± s.d. over 25 paired outer evaluations. Held-out models are evaluated on 2,232 molecules. ROC-AUC is computed from continuous predictions.

Evaluation	Model	R^2	RMSE	Spearman ρ	ROC-AUC
5×5 CV	XGB	0.237 ± 0.121	0.177 ± 0.018	0.551 ± 0.054	0.758 ± 0.034
	MLP	0.290 ± 0.114	0.171 ± 0.016	0.566 ± 0.057	0.764 ± 0.033
	GNN	$\mathbf{0.509 \pm 0.056}$	$\mathbf{0.142 \pm 0.010}$	$\mathbf{0.597 \pm 0.054}$	$\mathbf{0.771 \pm 0.032}$
Held-out	XGB	0.515	0.157	0.705	0.840
	MLP	0.591	0.144	**0.711**	0.840
	GNN	**0.615**	**0.140**	0.709	**0.850**

significant differences between all model pairs at a family-wise significance level of 0.05, supporting the ordering GNN > MLP > XGB (Figure 3a). We see the same ordering for RMSE, where the GNN is outperforming both MLP and XGB. There is no significant difference between GNN and MLP models on MAE and Spearman correlation instead. Parametric assumption tests and normality diagnostics are reported in Appendix D.1.

On the 2,232 held-out molecules, the GNN refit on the whole development set retained the highest R^2 (0.615) and lowest RMSE (0.140), followed by MLP (R^2 =0.591, RMSE=0.144) and XGB (R^2 =0.515, RMSE=0.157). Ranking performance was similar across models, with Spearman correlations ranging from 0.705 to 0.711. At the classification threshold of 0.7, the GNN achieved ROC-AUC 0.850, precision 0.65, and recall 0.53. XGB had slightly higher precision (0.66) but lower recall (0.20), while MLP achieved precision 0.58 and recall 0.39. More results are reported in Appendix D.2.

4.3 Ensemble Performance

Ensemble strategies were evaluated on 1,854 molecules, separately from the 463-molecule hill-climb selection subset (Table 2). The best single GNN achieved RMSE 0.1378 and $R^2 = 0.616$. Uniform averaging of all 75 models yielded RMSE 0.1346 and $R^2 = 0.633$, while averaging the 25 GNNs yielded 0.1332 and 0.641. The Caruana ensemble performed the best, with the reported composition of 13 GNN, seven MLP, and seven XGB members, achieving RMSE 0.1320 and $R^2 = 0.647$.

The best single model and Caruana ensemble had Spearman correlations of 0.68 and 0.72, respectively; see Figure 3b for parity plots. At the 0.7 threshold, precision was 0.51 and 0.66, while recall was 0.47 and 0.31. Both models assigned positive scores to many zero-labeled molecules, including predictions in the central score range.

Table 2: Ensemble results on the common 1,854-molecule evaluation subset, disjoint from the 463 cases (20%) used for Caruana selection (hill-climb set). Dispersion represents the mean disagreement among an ensemble's members, calculated by averaging the individual predictions' standard deviations. The value $\rho(|e|, s)$ is defined as the Spearman correlation between predictions' absolute errors and their corresponding standard deviations.

Strategy	n	Composition	RMSE	R^2	Dispersion	$\rho(\|e\|, s)$	Cov. $1\sigma/2\sigma/3\sigma$
Best single	1	GNN:1	0.1378	0.616	—	—	—
Uniform	75	XGB:25; MLP:25; GNN:25	0.1346	0.633	0.0693	0.332	46.9 / 70.3 / 83.7%
Best backend	25	GNN:25	0.1332	0.641	0.0481	**0.369**	38.2 / 66.6 / 82.0%
Greedy	27	XGB:7; MLP:7; GNN:13	**0.1320**	**0.647**	0.0675	0.337	47.0 / 72.9 / 84.4%

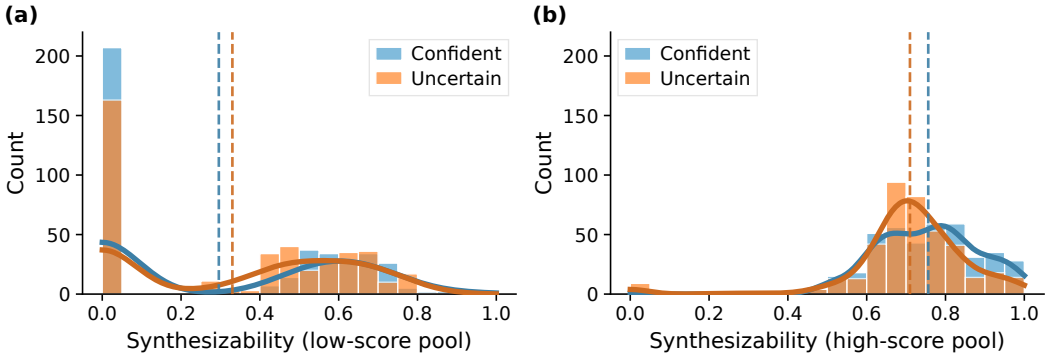


Figure 4: Route-derived synthesizability scores for the four candidate pools ($n = 400$ each) containing unseen PROTACs, grouped by RETROTAC ensemble confidence within the low-scored (a) and high-scored (b) tails. Dashed lines mark each subset's mean; curves are kernel density estimates (KDEs).

Mean member dispersion, a measure of how much ensemble members disagree with each other, was 0.0693 for uniform averaging, 0.0481 for the GNN average, and 0.0675 for Caruana selection; unsurprisingly, the members of the ensemble with the same backbone (GNN) agree most with each other. Spearman correlations between absolute error and member standard deviation were 0.332, 0.369, and 0.337, respectively, suggesting the potential to use ensemble standard deviation as a weakly calibrated error estimate.

4.4 PROSPECTIVE VALIDATION

We evaluated four subsets of 400 unseen PROTACs, stratified by predicted score and certainty. High-score candidates had predictions above 0.868, whereas low-score candidates had predictions below -0.058; each pool was divided into confident (high-certainty) and uncertain (low-certainty) subsets.[1]

Within the high-score pool (Figure 4b), mean route-derived synthesizability was 0.756 for confident candidates and 0.710 for uncertain candidates. Resolution rates were 99.75% (399/400) and 97.75% (391/400), respectively, with the latter including 390 fully solved routes and one directly purchasable molecule. Only one confident and nine uncertain candidates yielded no route. Structural summaries of the 399 confident and 390 uncertain non-empty trees showed short routes in both subsets: mean route depths were 2.17 and 2.35, and mean fragment balances were 0.633 and 0.585, respectively.

Within the low-score pool (Figure 4a), mean route-derived synthesizability was 0.296 for confident candidates and 0.330 for uncertain candidates. No route was found for 207/400 confident candidates (51.75%) and 163/400 uncertain candidates (40.75%), producing a pronounced concentration of scores at zero. Fully solved routes were obtained for 193/400 (48.25%) and 215/400 (53.75%), respectively. The uncertain subset also contained 22 partial routes, which were excluded from the

[1] RETROTAC predictions are not clipped to the $[0, 1]$ range of the training label, so the lowest-scoring candidates can fall slightly below zero.

resolution rate. Across the 193 confident and 237 uncertain non-empty trees, mean route depths were 3.45 and 3.94, and mean fragment balances were 0.471 and 0.409, respectively.

5 DISCUSSION

5.1 RELATIONSHIP TO EXISTING SYNTHESIZABILITY SCORES

The modest correlations in Section 4.1 indicate that molecule-level scores and our route-derived score capture overlapping but non-equivalent notions of accessibility. Score direction affects the signs but not the small magnitudes, since SA score and FSscore correlate far more strongly with each other ($\rho = -0.76$) than either does with our route-derived score.

The reason differs by family. SA score ($\rho = -0.37$) adds an unbounded heavy-atom penalty to its fragment term, so it registers a PROTAC's size rather than how few bimolecular reactions assemble it; the chemist ratings it was fitted to themselves correlate with atom count at $R^2 = 0.688$. SYBA ($\rho = 0.17$) avoids that size bias but still reads only fragment log-odds, with no account of how a molecule is built. SCScore ($\rho = -0.14$) is fitted along single reactions, ordering a product above its reactants but barely separating similarly sized PROTACs of differing route depth. The binary classifiers RAscore ($\rho = 0.25$) and GASA ($\rho = -0.27$) return one decision boundary and cannot rank molecules that all lie on its synthesizable side, as we do here. FSscore correlates most strongly ($\rho = 0.33$) but encodes a relative preference between molecules rather than a property of a route, and its PROTAC adaptation rests on 100 pairs from a single chemist and is not the default model.

Divergence alone does not make the route-derived score the better quantity to rank on; the prospective results do. Among unseen enumerated PROTACs (Section 4.4), the candidates RETROTAC ranked highest were resolved by the planner in 98.75% of cases against 51.00% for the lowest-ranked, at a mean route-derived synthesizability of 0.733 against 0.313, and gave shorter, more balanced routes. RETROTAC therefore offers two advantages the molecule-level scores lack. It discriminates among molecules that are all nominally synthesizable, and a low score is traceable, decomposing into route depth and fragment imbalance. Feasibility, selectivity and yield lie outside what it represents, and we make no claim that it supersedes scores validated against those endpoints.

5.2 IMPLICATIONS FOR CANDIDATE PRIORITIZATION

The ensemble's ranking performance supports its potential use for prioritizing molecules for subsequent route inspection. At the chosen threshold of 0.7, Caruana selection yields a more precise but less comprehensive selection than the best single model. The positive association between member disagreement and absolute error provides an additional signal that could help prioritize predictions for review.

On the enumerated PROTAC candidates, RETROTAC predicted successful route resolution and routes with fewer steps, shorter sequential depth, and more balanced fragment coupling. These structural differences are consistent with the intended preference of the route score for short, modular syntheses. The high-low score separation was larger in the confident pool, suggesting that ensemble disagreement may help identify candidates with more reliable extreme predictions. However, the uncertainty distributions overlapped substantially, and selecting score extremes separately within each pool does not isolate the contribution of uncertainty from differences in predicted score or molecular composition.

5.3 RETROSYNTHESIS PLANNING

Several properties of the planning tool influence the returned routes used for building the training set of route-derived synthesis scores in this work. Route quality depends primarily on the building-block collection, so the set of returned trees varies substantially across collections, and some collections are better suited to PROTAC synthesis than others; we found the eMolecules collection particularly appropriate here. Notably, however, the expansion policy does not decompose macrocycles, which accounts for a portion of the unresolved PROTACs. Because most molecules were resolved by a search depth of six, we do not expect deeper searches to raise the resolution rate. ShallowTree prioritizes speed and halts once it finds the first complete tree for a molecule, so it retains a single

route per target and may overlook better alternatives in exchange for throughput. Furthermore, a minority of routes terminate in building blocks that appear unusual from a synthetic-chemistry standpoint. All of them, however, are cataloged in the eMolecules building-block stock, so we treat them as valid, if initially surprising, starting materials rather than as artifacts of the search. These routes reflect what a retrosynthesis tool, here ShallowTree, proposes, and need not match how a chemist would assemble a PROTAC in practice, which often means preparing the warhead, linker, and E3 ligand separately and joining them through a coupling such as an amide bond formation (for example). We therefore read the scored routes as a first, planner-derived proxy for PROTAC synthesizability rather than as synthesis recipes. A natural extension is to assess the synthesizability of each component separately from that of the assembled PROTAC.

5.4 LIMITATIONS

The route-derived score summarizes a single route from one planner under a fixed stock and budget, and says nothing about experimental feasibility, selectivity, yield, or cost. Its endpoints are asymmetric, since a score of one can simply mean stock membership while a zero conflates real difficulty with a search that failed under the tested settings. The equal weighting of route depth and fragment balance is an uncalibrated design choice, and the 0.7 classification cutoff is a reporting convention, so we treat the continuous score and its ranking as the primary output. Validation against executed syntheses or expert route assessments is needed before reading the score as practical synthetic effort. Further, our search settings match common academic practice and resolved most PROTACs, so we do not attribute the unresolved fraction to under-resourcing. We keep one route per molecule, a standard simplification given that no planner enumerates the route space exhaustively; a resolved route is enough to show a molecule is reachable, which is what the label records. Sampling several routes per molecule, for instance to estimate a success probability, is left to future work. Additionally, the ensemble standard deviation tracks error only weakly and under-covers, so it cannot serve directly as a decision threshold; calibrating it on held-out data would make it more actionable. Finally, we have no experimentally validated negative examples, only enumerated PROTACs of unknown feasibility, so low scores and failed searches do not prove a molecule is hard or impossible to make. Prospective synthesis attempts with recorded failures would provide stronger validation, though structured data of this nature does not exist.

5.5 CONCLUSION

We set out to make a useful synthesizability surrogate for PROTAC assessment, where generative models now propose candidates faster than retrosynthesis models can evaluate them. We defined a transparent route score that rewards short route depth and the balanced joining of fragments, and we trained RETROTAC to predict it directly from a PROTAC structure, reaching a held-out RMSE of 0.13 and an R^2 of 0.65 as an ensemble. On more than one million enumerated PROTACs, the candidates RETROTAC ranked highest were resolved by the planner far more often than its lowest-ranked (98.75% versus 51.00%) and yielded shorter, more balanced routes, so the surrogate reserves explicit planning for the most promising molecules. We release nearly 20,000 retrosynthetic routes and the trained ensemble to support further work into the computer-aided synthesis planning of large, complex molecules.

RETROTAC captures route depth and modular balance rather than experimental feasibility, and validation against executed syntheses remains the natural next step. Even so, it offers a fast and interpretable way to connect virtual screening with the route characteristics that make a PROTAC practical to build.

AI USE STATEMENT

Generative AI tools were used to edit manuscript prose, propose organization and figure designs, identify potentially relevant literature, and provide critical feedback on statistical reporting and experimental design. They were not used to generate the reported experimental data or numerical model results. The authors reviewed the AI-assisted material, will verify every citation against the primary source, and take responsibility for the final text, claims, code and analyses.

Ethics statement

This study uses molecular structures and computationally generated route descriptors and does not involve human participants, personal data or animal experiments. Synthetic-accessibility predictions can be over-interpreted as guarantees of laboratory feasibility. We mitigate this risk by defining the target explicitly, reporting its planner dependence and residual calibration, and requiring retrosynthetic and expert review before experimental decisions. Data, software and model releases will follow the licences and access restrictions of their upstream sources.

Reproducibility statement

The experimental pipeline uses fixed data assignments, five recorded partition seeds, containerized environments and cached molecular features. The appendix specifies the target, grouping, optimization, metrics and ensemble algorithms. An anonymous supplementary archive containing source code, environment/container recipes, configuration files, split assignments and commands for dataset assembly, models training and evaluation, and plotting can be found here: `https://anonymous.4open.science/r/RetroTAC-8C28/`.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

Jeremy R. Ash, Cas Wognum, Raquel Rodríguez-Pérez, Matteo Aldeghi, Alan C. Cheng, Djork-Arné Clevert, Ola Engkvist, Cheng Fang, Daniel J. Price, Jacqueline M. Hughes-Oliver, and W. Patrick Walters. Practically Significant Method Comparison Protocols for Machine Learning in Small Molecule Drug Discovery. *Journal of Chemical Information and Modeling*, 65(18):9398–9411, September 2025. doi: 10.1021/acs.jcim.5c01609.

Guy W. Bemis and Mark A. Murcko. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

María J. Blanca, Rafael Alarcón, Jaume Arnau, Roser Bono, and Rebecca Bendayan. Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*, 50 (3):937–962, June 2018. ISSN 1554-3528. doi: 10.3758/s13428-017-0918-2. URL `https://doi.org/10.3758/s13428-017-0918-2`.

Jackson W Burns, Akshat Shirish Zalte, Charlles RA Abreu, Jochen Sieg, Christian Feldmann, Miriam Mathea, and William H Green. Deep Learning Foundation Models from Classical Molecular Descriptors. *arXiv preprint arXiv:2506.15792*, 2025.

Darko Butina. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 18, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015432. URL `https://doi.org/10.1145/1015330.1015432`.

Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: learning retrosynthetic planning with neural guided a* search. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Shuan Chen and Yousung Jung. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *Journal of Cheminformatics*, 2024. doi: 10.1186/s13321-024-00879-0.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Connor W. Coley, Luke Rogers, William H. Green, and Klavs F. Jensen. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of Chemical Information and Modeling*, 58 (2):252–261, 2018. doi: 10.1021/acs.jcim.7b00622.

eMolecules. eMolecules, 2026. URL https://www.emolecules.com.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009. doi: 10.1186/1758-2946-1-8.

Sinuo Gao, Xiaofei Zhou, Lu Liang, and Jianping Lin. RetroScore: graph edit distance-guided retrosynthesis for accessibility scoring with route metrics. *Journal of Cheminformatics*, 18(1):10, 2025.

Jingxuan Ge, Shimeng Li, Gaoqi Weng, Huating Wang, Meijing Fang, Huiyong Sun, Yafeng Deng, Chang-Yu Hsieh, Dan Li, and Tingjun Hou. PROTAC-DB 3.0: an updated database of PROTACs with extended pharmacokinetic parameters. *Nucleic Acids Research*, 53(D1):D1510–D1515, 2025.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics*, 12(1):70, November 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00472-1. URL https://doi.org/10.1186/s13321-020-00472-1.

Yossra Gharbi and Rocío Mercado. A comprehensive review of emerging approaches in machine learning for de novo PROTAC design. *Digital Discovery*, 3(11):2158–2176, 2024.

David E Graff, Nathan K Morgan, Jackson W Burns, Anna C Doner, Brian Li, Shih-Cheng Li, Joel Manu, Angiras Menon, Hao-Wei Pang, Haoyang Wu, et al. Chemprop v2: An efficient, modular machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 66(1):28–33, 2026.

Jeff Guo, Franziska Knuth, Christian Margreitter, Jon Paul Janet, Kostas Papadopoulos, Ola Engkvist, and Atanas Patronov. Link-INVENT: generative linker design with reinforcement learning. *Digital Discovery*, 2(2):392–408, 2023. doi: 10.1039/D2DD00115B. URL http://dx.doi.org/10.1039/D2DD00115B.

Steffen Herbold. Autorank: A Python Package for Automated Ranking of Classifiers. *Journal of Open Source Software*, 5(48):2173, 2020. doi: 10.21105/joss.02173.

Zhenyi Hu and Craig M. Crews. Recent Developments in PROTAC-Mediated Protein Degradation: From Bench to Clinic. *ChemBioChem*, 23(2):e202100270, 2022. doi: https://doi.org/10.1002/cbic.202100270. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cbic.202100270.

David Kreutter and Jean-Louis Reymond. Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search. *Chemical Science*, 14(36): 9959–9969, 2023.

Baiqing Li, Ting Ran, and Hongming Chen. 3D based generative PROTAC linker design with reinforcement learning. *Briefings in Bioinformatics*, 24(5):bbad323, September 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad323. URL https://doi.org/10.1093/bib/bbad323.

Junren Li, Lei Fang, and Jian-Guang Lou. Retro-BLEU: quantifying chemical plausibility of retrosynthesis routes through reaction template sequence analysis. *Digital Discovery*, 3(3):482–490, 2024. ISSN 2635-098X. doi: https://doi.org/10.1039/d3dd00219e. URL https://www.sciencedirect.com/science/article/pii/S2635098X24000433.

Alan H Lipkus. A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry*, 26(1):263–265, 1999.

Nir London and Jaime Prilusky. PROTACpedia, 2024. URL https://protacpedia.weizmann.ac.il.

Krzysztof Maziarz, Austin Tripp, Guoqing Liu, Megan Stanley, Shufang Xie, Piotr Gaiński, Philipp Seidl, and Marwin H. S. Segler. Re-evaluating retrosynthesis algorithms with Syntheseus. *Faraday Discussions*, 256:568–586, January 2025. ISSN 1359-6640. doi: 10.1039/d4fd00093e. URL https://doi.org/10.1039/d4fd00093e.

Rebecca M. Neeser, Bruno Correia, and Philippe Schwaller. FSscore: A Personalized Machine Learning-Based Synthetic Feasibility Score. *Chemistry–Methods*, 4(11):e202400024, 2024. doi: 10.1002/cmtd.202400024.

Divya Nori, Connor W. Coley, and Rocío Mercado. De novo PROTAC design using graph-based deep generative models. In *Proceedings of the NeurIPS 2022 Workshop on AI for Science*, 2022. doi: 10.48550/arXiv.2211.02660. URL https://openreview.net/forum?id=pGyp4o9gky0.

Edward Price, Manuel Weinheimer, Alexey Rivkin, Gary Jenkins, Marjoleen Nijsen, Philip B Cox, and David DeGoey. Beyond rule of five and PROTACs in modern drug discovery: polarity reducers, chameleonicity, and the evolving physicochemical landscape. *Journal of Medicinal Chemistry*, 67 (7):5683–5698, 2024.

Xinran Qin, Yinpeng Zhang, Yajunzi Wang, Yintao Zhang, Jiachen Jing, Yuyuan Zhang, Gaoxiang Xu, Haoping Teng, Tianjun Wang, Lei Fu, and others. TPDdb: the comprehensive database of targeted protein degrader. *Nucleic Acids Research*, 54(D1):D1683–D1691, 2026.

Stefano Ribes, Ranxuan Zhang, Télio Cropsal, Anders Källberg, Christian Tyrchan, Eva Nittinger, and Rocío Mercado. PROTAC-Splitter: a machine learning framework for automated identification of PROTAC substructures. *Journal of Cheminformatics*, 18(1):30, February 2026. ISSN 1758-2946. doi: 10.1186/s13321-025-01135-9. URL https://doi.org/10.1186/s13321-025-01135-9.

Xiangyan Sun, Ke Liu, Yuquan Lin, Lingjie Wu, Haoming Xing, Minghong Gao, Ji Liu, Suocheng Tan, Zekun Ni, Qi Han, Junqiu Wu, and Jie Fan. ChemiRise: a data-driven retrosynthesis engine, 2021. URL https://arxiv.org/abs/2108.04682.

Amol Thakkar, Thierry Kogej, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.*, 11(1):154–168, 2020. doi: 10.1039/C9SC04944D. URL http://dx.doi.org/10.1039/C9SC04944D.

Amol Thakkar, Veronika Chadimová, Esben Jannik Bjerrum, Ola Engkvist, and Jean-Louis Reymond. Retrosynthetic Accessibility Score (RAscore) – Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning. *Chemical Science*, 12(9):3339–3349, 2021. doi: 10.1039/D0SC05401A.

Milan Voršilák, Michal Kolář, Ivan Čmelo, and Daniel Svozil. SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds. *Journal of Cheminformatics*, 12(1):35, 2020. doi: 10.1186/s13321-020-00439-2.

Dhanushka Weerakoon, Rodrigo J. Carbajo, Leonardo De Maria, Christian Tyrchan, and Hongtao Zhao. Impact of PROTAC Linker Plasticity on the Solution Conformations and Dissociation of the Ternary Complex. *Journal of Chemical Information and Modeling*, 62(2):340–349, January 2022. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c01036. URL https://doi.org/10.1021/acs.jcim.1c01036.

Gaoqi Weng, Xuanyan Cai, Dongsheng Cao, Hongyan Du, Chao Shen, Yafeng Deng, Qiaojun He, Bo Yang, Dan Li, and Tingjun Hou. PROTAC-DB 2.0: an updated database of PROTACs. *Nucleic Acids Research*, 51(D1):D1367–D1372, 2023.

Mingyuan Xu, Chaoming Huang, Li Pang, Qirui Deng, Hao Zhang, Anjie Qiao, Zhiwen Luo, Zhen Wang, Chang-Yu Hsieh, Zhang Zhang, Tie-Gen Chen, Hongming Chen, and Jinping Lei. Design of Synthesizable PROTACs through Synthesis Constrained Generative Model and Reinforcement Learning. *JACS Au*, 6(7):4239–4252, July 2026. ISSN 2691-3704. doi: 10.1021/jacsau.6c00732. URL https://doi.org/10.1021/jacsau.6c00732.

Ran Zhang, Shihang Wang, Lin Wang, Siyuan Tian, Yilin Tang, and Fang Bai. DeepPSA: A Geometric Deep Learning Model for PROTAC Synthetic Accessibility Prediction. *Journal of Chemical Information and Modeling*, 65(13):6861–6873, June 2025. ISSN 1549-9596. doi: 10.1021/acs.jcim.5c00366. URL https://doi.org/10.1021/acs.jcim.5c00366.

A Data distributions

A.1 Held-out clustering

We evaluated Butina Tanimoto-distance cutoffs from 0.50 to 0.95 in increments of 0.05 using 2,048-bit, radius-8 Morgan fingerprints of all 18,076 scored PROTACs. From 0.50 to 0.85, silhouette and Calinski–Harabasz scores improved, whereas the Davies–Bouldin index increased, indicating poorer quality by that criterion. We combined the three indices by min–max normalizing each, inverting Davies–Bouldin so that higher values consistently indicate better quality, and taking their mean (Figure 5).

The cutoff also affected the held-out fraction because selection operated on complete clusters. The selection procedure returned approximately 10% of molecules at cutoffs up to 0.70, 12–15% at 0.75–0.85, and 100% at 0.90 and 0.95. At 0.90, a single cluster contained 18,066 of the 18,076 molecules. We therefore restricted selection to cutoffs yielding a held-out fraction within three percentage points of the 10% target and chose the highest composite quality among them. The selected cutoff, 0.85, produced 49 clusters in total and a held-out set of 2,232 molecules (12.3%).

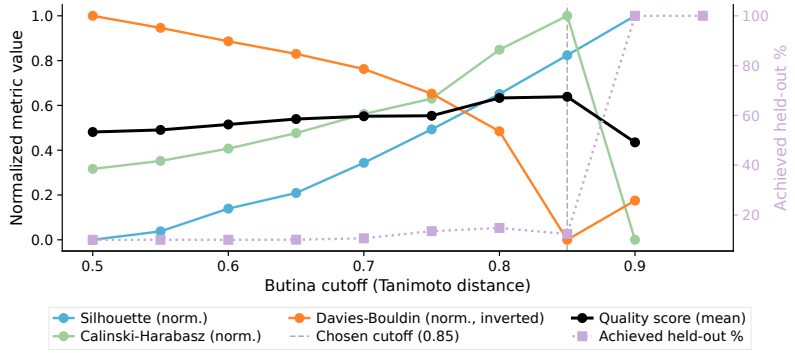


Figure 5: Cluster quality against Butina cutoff. Silhouette and Calinski–Harabasz are min-max normalized, Davies–Bouldin normalized and inverted, and the black curve is their mean. The right axis gives the attainable held-out fraction. The dashed rule marks the adopted cutoff of 0.85.

A.2 Cross-validation fold distributions

We reconstructed the 25 outer validation folds using the grouping function and seeds used for training. The 15,844 development molecules comprised 2,327 warhead scaffold groups, the largest containing 609 molecules (3.8%). Despite this variation in group size, greedy assignment produced folds of similar size: 3,169±46 molecules and 465±62 groups, with every fold within 5% of an equal split (Figure 6a).

Label distributions were also consistent across folds. All fold medians were within 0.03 of the development-set median, which lay within every fold's interquartile range (Figure 6b). Fold means ranged from 0.596 to 0.648 around the development mean of 0.621, while the proportions at $y = 0$ and $y = 1$ were 7.4±1.3% and 1.9±0.4%, respectively. The empirical cumulative distribution functions closely followed that of the development set (Figure 6c), with a mean Kolmogorov–Smirnov statistic of 0.054. The held-out distribution differed more substantially (0.181), including a larger proportion of zero scores.

B Component leakage and cross-validation grouping

We evaluated twelve grouping keys on the 15,844 development-set PROTACs: the whole molecule, each component, and their pairwise and three-way combinations. We used generic Bemis–Murcko molecular scaffolds as grouping keys (Bemis & Murcko, 1996), with exact-SMILES fallback for acyclic inputs. Composite keys grouped molecules either by joint component identity or by connected components linking molecules that shared any constituent key. Component overlap was measured

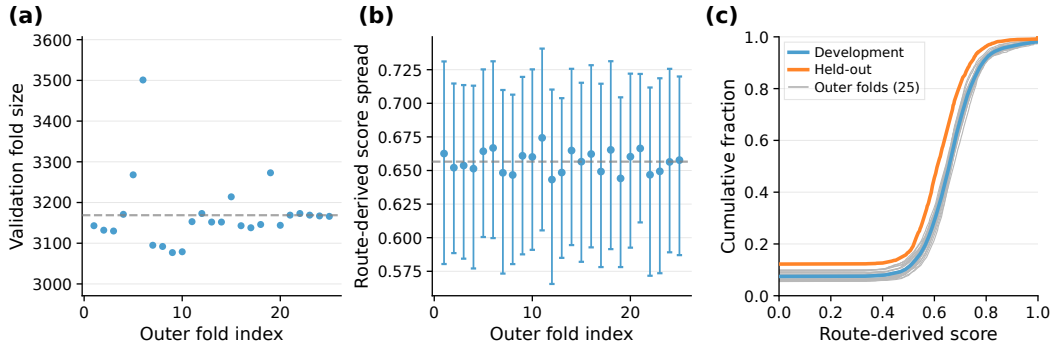


Figure 6: Distributions across the 25 outer validation folds (five seeds × five folds). (a) Fold sizes, with an equal split ($n/5 \approx 3,169$) shown by the dashed line. (b) Fold medians and interquartile ranges of the route-derived score; the dashed line marks the development-set median. (c) Empirical cumulative distribution functions for the validation folds (grey), development set (blue) and held-out set (orange).

Table 3: Component overlap for representative grouping keys, reported as mean ± s.d. (%) across 25 outer CV splits. Scaffold overlap uses generic Bemis–Murcko molecular scaffolds; exact-SMILES overlap measures matching component strings. Exact full-PROTAC SMILES overlap was zero under every grouping strategy. "All components" denotes connected-component grouping across warhead, linker, and E3 ligand keys; its largest group contains 99.9% of the dataset. The adopted grouping key is shown in bold.

	Scaffold overlap (%)			Exact-SMILES overlap (%)		
Grouped by	Warhead	Linker	E3 ligand	Warhead	Linker	E3 ligand
Whole molecule	84.6 ± 0.6	83.8 ± 0.8	93.4 ± 0.6	79.2 ± 0.7	71.8 ± 0.8	89.2 ± 0.9
Warhead	0.0 ± 0.0	77.9 ± 3.3	91.6 ± 2.6	0.0 ± 0.0	64.0 ± 3.7	86.8 ± 2.9
Linker	81.4 ± 2.8	0.0 ± 0.0	92.4 ± 3.0	75.7 ± 3.1	0.0 ± 0.0	87.7 ± 4.1
E3 ligand	76.5 ± 5.1	79.0 ± 5.3	0.0 ± 0.0	70.2 ± 5.1	64.1 ± 6.2	0.0 ± 0.0
All components	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

across the same five seeds and five outer folds used during training. No identical full-PROTAC SMILES occurred in both training and validation under any grouping strategy; the overlap reported here concerns individual components.

Whole-molecule grouping separated complete molecular representations but retained substantial component overlap (Table 3): 79.2% of validation warheads, 71.8% of linkers, and 89.2% of E3 ligands also occurred in training as exact SMILES matches. When two or three component types were required to be disjoint across folds, molecules linked directly or indirectly through any shared component key had to remain in the same group. The largest resulting group contained 93.2–99.9% of the dataset, precluding approximately balanced five-fold CV. Although whole-molecule groups had a higher association with the route-derived score than warhead groups ($\eta^2 = 0.884$ versus 0.575), they formed a much finer partition: 11,686 groups, with 59.2% of molecules in singleton groups, compared with 3,128 groups and 10.6% under warhead grouping. Here, η^2 measures the fraction of observed score variance attributable to differences between group means; its dependence on partition granularity prevents interpreting a higher value as stronger protection against component leakage.

We therefore selected warhead grouping: it eliminated warhead overlap, retained manageable group sizes (largest group: 434 molecules, or 2.7%), and had the highest observed association with the route-derived score among individual components. This grouping also evaluates generalization to unseen warhead groups while allowing linkers and E3 ligands already encountered during training.

Table 4: Candidate score components across 16,318 scorable routes. Tied is the percentage sharing the most common score; effective levels is the exponential of the Shannon entropy of the score distribution. R^2 from adopted measures how well a component is explained by regression on route depth and balance. The adopted components are bold.

Term	Mean	s.d.	IQR	Tied (%)	Eff. levels	R^2 from adopted	ρ with y
Depth D	0.752	0.124	0.100	33.4	5.0	—	0.723
Balance B	0.568	0.126	0.160	0.1	11,995.3	—	0.755
Joining fraction	0.896	0.124	0.200	52.7	7.1	0.260	0.305
Convergence	0.291	0.185	0.229	21.9	11.9	0.194	−0.297
Branching (rescaled)	0.453	0.065	0.100	53.1	7.5	0.203	0.265

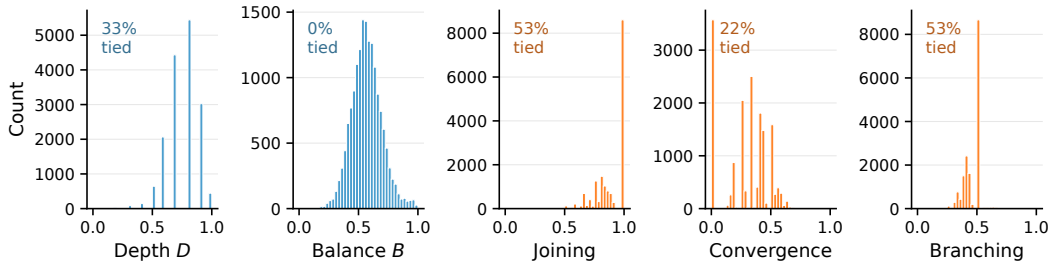


Figure 7: Distributions of the five candidate score components across 16,318 scorable routes. Blue denotes the adopted components; orange denotes the alternatives. Tied is the percentage of routes sharing each component's most common value.

C Score design and analysis

We examined the choice of score components and the sensitivity of the adopted two-term route-derived score to its weights, saturation depth and length-decay function. Of the 18,076 scored PROTACs, 16,318 had scorable routes; the remaining 304 purchasable and 1,454 unsolved targets retained fixed scores of 1 and 0 under every variant. The analyses below exclude these fixed endpoints. Including them increased correlations by 0.02–0.11 without changing the qualitative conclusions.

Candidate components. Fragment balance provided the finest resolution, with 13,375 distinct values and no value shared by more than 0.14% of routes. The depth term was coarser, with five effective levels (Table 4, Figure 7). Joining fraction and rescaled branching assigned their most common value to 52.7% and 53.1% of routes, respectively; branching also had approximately half the standard deviation of either adopted term. Convergence was negatively correlated with the adopted score ($\rho = -0.297$), consistent with deeper routes accumulating more off-path steps. These distributions support using depth and balance to distinguish routes in this dataset: joining is already common, whereas the relative sizes of the joined fragments vary substantially.

Sensitivity to parameter choices. Rank agreement remained high when the two adopted components were retained, with Spearman correlations of 0.906–0.997 across the tested parameter settings (Table 5). Changing the saturation depth from $D_{\max} = 11$ to 9 or 13 had little effect on ranking ($\rho \geq 0.995$); larger changes to 7 or 15 retained $\rho \geq 0.974$. Unequal length/balance weights of 0.25/0.75 and 0.75/0.25 had a greater effect ($\rho = 0.917$ and 0.906). Logarithmic length decay largely preserved ranking ($\rho = 0.989$) but shifted the score distribution, with a mean absolute change of 0.122 and a maximum of 0.139. At the fixed classification threshold of 0.7, this shift changed 25.1% of labels, showing that rank agreement does not imply stability of thresholded classifications.

Sensitivity to component selection. Changing the component set generally had a larger effect on ranking. Adding branching with equal weights retained the highest agreement ($\rho = 0.944$), whereas adding joining at weights of 0.50/0.30/0.20 reduced agreement to 0.791; including all five components gave 0.595 (Table 5). Component selection therefore remains a substantive design choice: the adopted

Table 5: Agreement between score variants and the adopted definition across 16,318 scorable routes. Label flips is the percentage whose binary label changes at a threshold of 0.7. Top-10% J is the Jaccard overlap between top-decile selections. Two-term weights are listed as length/balance; the three-term weights are length/balance/joining.

Variant	Spearman ρ	Kendall τ	Mean $\lvert\Delta\rvert$	Label flips (%)	Top-10% J
Same two terms, re-parameterized					
Saturation $D_{\max}$ = 7	0.974	0.878	0.082	16.6	0.788
Saturation $D_{\max}$ = 9	0.995	0.945	0.031	8.4	0.908
Saturation $D_{\max}$ = 13	0.997	0.954	0.021	7.5	0.921
Saturation $D_{\max}$ = 15	0.989	0.917	0.035	14.5	0.851
Logarithmic depth decay	0.989	0.932	0.122	25.1	0.784
Weights 0.25/0.75	0.917	0.762	0.050	13.4	0.657
Weights 0.75/0.25	0.906	0.779	0.050	23.0	0.626
Different term set					
Depth only	0.723	0.588	0.100	50.1	0.289
Balance only	0.755	0.579	0.100	19.2	0.496
All five terms, equal weights	0.595	0.435	0.080	31.1	0.259
+ branching, equal weights	0.944	0.822	0.070	25.9	0.764
+ joining, equal weights	0.853	0.716	0.082	37.9	0.695
+ joining, 0.50/0.30/0.20	0.791	0.627	0.101	43.5	0.562
+ convergence, equal weights	0.576	0.419	0.125	31.6	0.255

score prioritizes route depth and fragment-size balance, while the alternatives can substantially alter rankings.

Implementation details. Fragment balance is averaged over bimolecular steps. Routes without such steps receive $B = 0.5$, so their scores vary only with depth. This fallback is uncommon in the present dataset but could be more frequent in datasets with fewer fragment-joining reactions.

C.1 Models and hyperparameter optimization

XGB. XGBoost uses 1,024-bit Morgan fingerprints (radius 8, chirality enabled) concatenated with 2D RDKit descriptors. Hyperparameter optimization determines whether to apply truncated SVD to the fingerprints and the number of retained components. Descriptors are median-imputed, constant columns are removed, and the remaining columns are standardized. The target is quantile-transformed, and training uses the pseudo-Huber loss. All preprocessing is fitted only on the relevant training partition.

MLP. The PyTorch Lightning multilayer perceptron uses the same representation and preprocessing as XGBoost, with smooth L1 loss, a batch size of 256, a maximum of 500 epochs and early-stopping patience of 20 epochs.

GNN. The Chemprop directed message-passing network is initialized from CheMeleon, which was pretrained to predict Mordred descriptors from approximately one million PubChem structures (Graff et al., 2026; Burns et al., 2025). It receives molecular graphs parsed from SMILES and is fine-tuned end-to-end using mean squared error.

For each outer fold, Optuna's TPE sampler evaluates up to 30 XGB, 25 MLP and 15 GNN trials over the search space in Table 6 (Akiba et al., 2019). Sampler seeds depend on the outer seed and fold index. A median pruner uses five startup trials and three warm-up steps; only the MLP reports intermediate objectives and is therefore eligible for pruning. Hyperparameters are selected on a grouped 80/20 split of the outer-training set by minimising

$$\mathcal{L} = 0.5\,\mathrm{RMSE} + 0.5\frac{1-\rho_s}{2}, \tag{2}$$

where ρ_s is Spearman correlation. The selected configuration is refitted on the outer-training set and evaluated once on its validation fold. Final models are fitted on the full development set; the held-out set is excluded from hyperparameter selection.

Table 6: Hyperparameter search space for the reported experiments. Log denotes logarithmic sampling; zero SVD components disables dimensionality reduction.

Backend	Parameter	Type	Range
XGB	n_estimators	int	[200, 3000]
	max_depth	int	[3, 8]
	learning_rate	float, log	$[10^{-3},\ 0.3]$
	subsample	float	[0.6, 1.0]
	colsample_bytree	float	[0.6, 1.0]
	min_child_weight	int	[1, 10]
	gamma	float	[0, 5]
	svd_components	int	[0, 128]
MLP	n_layers	int	[1, 4]
	hidden_dim	categorical	{64, 128, 256, 512}
	dropout	float	[0, 0.5]
	learning_rate	float, log	$[10^{-4},\ 10^{-2}]$
	weight_decay	float, log	$[10^{-6},\ 10^{-2}]$
	svd_components	int	[0, 128]
GNN	ffn_hidden_dim	categorical	{128, 256, 512}
	ffn_n_layers	int	[1, 3]
	dropout	float	[0, 0.4]
	max_lr	float, log	$[10^{-4},\ 5 \times 10^{-3}]$

Table 7: Selected hyperparameters across 25 outer folds and the final-model configurations. Continuous parameters and larger integer ranges are summarized as median [interquartile range]; categorical parameters and small integer ranges are reported as mode (frequency).

Backend	Parameter	Across the 25 folds	Final model
XGB	n_estimators	1,762 [1,280, 2,279]	2,037
	max_depth	7 (9/25)	7
	learning_rate	1.2×10^{-2} [5.0×10^{-3}, 1.9×10^{-2}]	4.7×10^{-3}
	subsample	0.859 [0.718, 0.907]	0.822
	colsample_bytree	0.727 [0.658, 0.858]	0.784
	min_child_weight	4 (5/25)	9
	gamma	2.30 [0.80, 3.14]	2.30
	svd_components	62 [41, 92]	116
MLP	n_layers	4 (16/25)	4
	hidden_dim	256 (8/25)	128
	dropout	0.305 [0.214, 0.412]	0.189
	learning_rate	2.6×10^{-3} [1.2×10^{-3}, 4.3×10^{-3}]	5.0×10^{-3}
	weight_decay	2.6×10^{-4} [3.2×10^{-5}, 1.4×10^{-3}]	7.9×10^{-5}
	svd_components	68 [36, 105]	36
GNN	ffn_hidden_dim	512 (10/25)	256
	ffn_n_layers	2 (11/25)	1
	dropout	0.223 [0.124, 0.305]	0.298
	max_lr	6.7×10^{-4} [3.3×10^{-4}, 1.4×10^{-3}]	6.6×10^{-4}

Across the 25 outer folds, the search comprised 750 XGB, 625 MLP and 375 GNN trials; 52.5% of MLP trials were pruned. By trial ten, all three models were within 2% of their final best objective, with no subsequent improvement for the GNN in the reported trajectory (Figure 8a). Final objectives were 0.167 ± 0.010 for GNN, 0.188 ± 0.011 for MLP and 0.201 ± 0.012 for XGB. The GNN achieved both lower RMSE and higher Spearman correlation during tuning (Figure 8b).

Selected hyperparameters varied across folds (Table 7). For example, the XGB SVD dimension had an interquartile range of 41–92, and the most common MLP and GNN hidden dimensions were selected in fewer than half the folds. We therefore report their distributions alongside the final-model configurations rather than summarizing each search by a single setting.

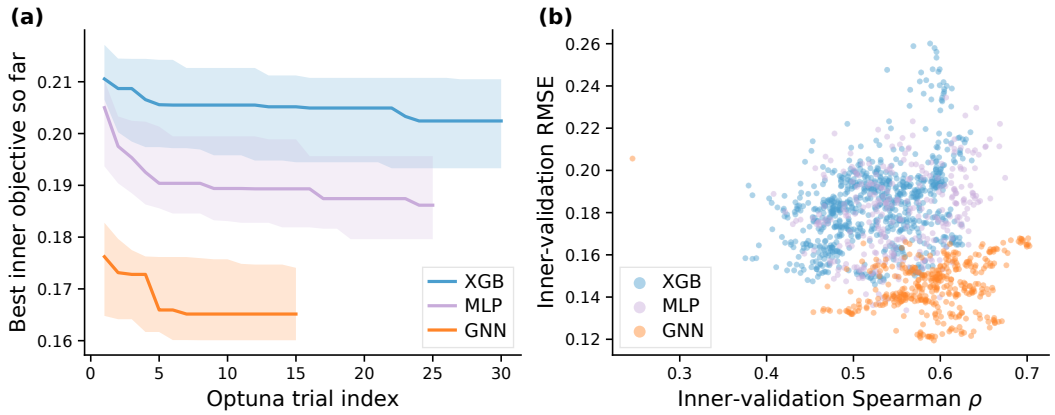


Figure 8: Hyperparameter search progress. (a) Best inner-validation objective by trial number, shown as the median and interquartile range across 25 outer folds. Traces end at each model's trial budget; pruned trials count towards the budget but cannot update the best objective. (b) Inner-validation RMSE and Spearman correlation for all completed trials.

Table 8: Between-model variance diagnostics for 25 paired outer-fold evaluations. The stated decision rule selects the parametric procedure unless both Levene's $p < 0.05$ and the maximum-to-minimum variance ratio exceeds 9.

Metric	Levene W	Levene p	Max/min variance	Branch
R^2	2.518	0.088	4.68	parametric (forced)
RMSE	2.900	0.062	3.21	parametric (forced)
MAE	1.702	0.190	2.39	parametric (forced)
Spearman ρ	0.021	0.980	1.10	parametric (forced)

D Additional performance results

D.1 Normality and variance diagnostics

Following the Polaris method-comparison protocol (Ash et al., 2025), we assessed the suitability of parametric comparisons using variance diagnostics and the shape of the fold-score distributions. The protocol motivates ANOVA and Tukey HSD by their robustness to moderate departures from normality in balanced comparisons, with each model contributing 25 evaluations from 5×5 repeated cross-validation. Accordingly, we examined Q–Q plots and reported Shapiro–Wilk tests as supplementary diagnostics, rather than using normality-test rejection alone to select the testing procedure.

For variance homogeneity, we adopted the protocol's maximum-to-minimum variance-ratio threshold of 9 for equally sized groups (Blanca et al., 2018). Our decision rule in Section 3.4 flags substantial heteroscedasticity only when this threshold is exceeded and Levene's $p < 0.05$. None of the four metrics met this criterion (Table 8): variance ratios ranged from 1.10 to 4.68 and Levene's p-values from 0.062 to 0.980. For R^2, the variance ratio was 4.6849 and Levene's $p = 0.0877$.

Shapiro–Wilk tests rejected normality at $\alpha = 0.05$ for two of the twelve model–metric combinations, both involving the MLP (Table 9, Figure 9). The R^2 rejection ($p = 0.001$) was associated with one unusually low result for one particular fold, $R^2 = -0.12$, approximately 3.6 standard deviations below the MLP mean. Excluding this observation for diagnostic purposes increased the Shapiro–Wilk p-value to 0.87. The MAE distribution also departed from normality ($p = 0.021$), with negative excess kurtosis; removing its largest value did not eliminate the rejection ($p = 0.026$). All 25 observations were retained in the statistical comparisons. We used repeated-measures ANOVA followed by Tukey HSD in accordance with the stated protocol, while documenting here these departures from normality.

Table 9: Shapiro–Wilk tests, sample skewness and excess kurtosis for the 25 outer-fold scores of each model. Rejections at $\alpha = 0.05$ are bold.

Model	Metric	Skew	Excess kurtosis	W	p
XGB	R^2	−1.204	2.182	0.925	0.068
	RMSE	0.674	0.004	0.942	0.166
	MAE	0.726	−0.180	0.933	0.104
	Spearman ρ	0.200	−0.735	0.975	0.761
MLP	R^2	−1.922	6.521	**0.842**	**0.001**
	RMSE	0.302	−1.153	0.932	0.098
	MAE	0.417	−1.240	**0.903**	**0.021**
	Spearman ρ	−0.143	−0.053	0.985	0.963
GNN	R^2	0.682	0.472	0.954	0.313
	RMSE	0.252	−1.286	0.928	0.080
	MAE	0.446	−0.118	0.953	0.295
	Spearman ρ	0.111	−1.145	0.956	0.347

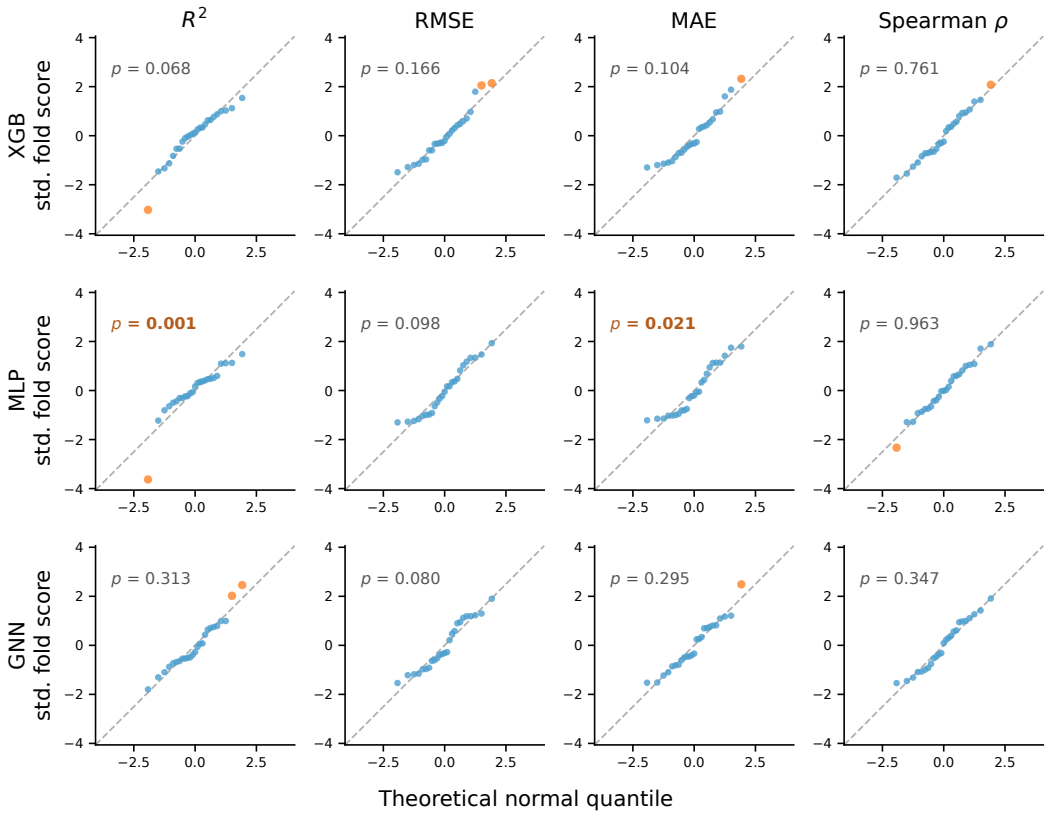


Figure 9: Normal Q–Q plots of the 25 outer-fold scores for each model and metric. Scores are standardized within panels, which share a common scale and the reference line $y = x$. Orange points lie more than two standard deviations from the mean. Annotations give Shapiro–Wilk p-values, with rejections at $\alpha = 0.05$ in bold.

D.2 Additional regression and classification metrics

Tables 10 and 11 report metrics that complement the main-text results; Figure 10 shows their fold distributions. In cross-validation, mean Pearson and Kendall correlations, explained variance, PR-AUC, MCC, F_1, balanced accuracy and Cohen's κ all followed the ordering GNN > MLP > XGB.

Table 10: MAE, precision and recall. CV values are mean ± s.d. over 25 paired outer evaluations; held-out values are evaluated on 2,232 molecules. Precision and recall use a threshold of 0.7 for both reference scores and predictions. The best value within each evaluation setting is bold.

Evaluation	Model	MAE	Precision	Recall
5×5 CV	XGB	0.104±0.010	**0.63±0.06**	0.31±0.10
	MLP	0.099±0.009	0.61±0.04	0.46±0.08
	GNN	**0.094±0.007**	0.59±0.04	**0.53±0.12**
Held-out	XGB	0.085	**0.66**	0.20
	MLP	**0.076**	0.58	0.39
	GNN	0.086	0.65	**0.53**

Table 11: Additional regression, ranking and classification metrics. CV values are mean ± s.d. over 25 paired outer evaluations. Held-out values are from models refitted on the full development set and evaluated on 2,232 molecules. Classification metrics use a threshold of 0.7; PR-AUC uses continuous predictions against the thresholded reference scores. The best value within each evaluation setting is bold, except for bias and predicted-positive rate.

	5×5 CV			Held-out		
Metric	XGB	MLP	GNN	XGB	MLP	GNN
MedAE	0.056±0.003	**0.054±0.003**	0.061±0.004	0.043	**0.042**	0.052
Max error	0.928±0.114	0.979±0.069	**0.819±0.123**	**0.704**	0.819	0.704
Expl. variance	0.250±0.116	0.298±0.105	**0.512±0.054**	0.516	0.592	**0.616**
Bias	-0.020±0.013	-0.013±0.012	-0.001±0.011	−0.007	0.010	−0.007
Pearson r	0.577±0.056	0.619±0.053	**0.717±0.037**	0.740	0.778	**0.792**
Kendall τ	0.393±0.043	0.408±0.045	**0.432±0.044**	0.523	**0.533**	0.532
PR-AUC	0.574±0.051	0.588±0.045	**0.602±0.050**	0.551	0.538	**0.574**
MCC	0.280±0.075	0.338±0.064	**0.353±0.066**	0.301	0.389	**0.505**
F_1	0.410±0.105	0.519±0.062	**0.548±0.081**	0.309	0.470	**0.582**
Specificity	**0.908±0.032**	0.852±0.041	0.813±0.064	**0.977**	0.937	0.937
Balanced acc.	0.611±0.041	0.655±0.034	**0.670±0.040**	0.590	0.666	**0.732**
Accuracy	0.710±0.036	**0.722±0.036**	0.720±0.033	0.836	0.838	**0.863**
Cohen's κ	0.251±0.086	0.329±0.067	**0.347±0.072**	0.245	0.379	**0.501**
Pred. pos. rate	0.166±0.055	0.251±0.055	0.300±0.083	0.056	0.123	0.147

Several metrics highlighted different aspects of performance. Specificity was highest for XGB (0.908), followed by MLP (0.852) and GNN (0.813), consistent with XGB's higher precision and lower recall. Overall accuracy was similar for MLP and GNN (0.722 versus 0.720), whereas balanced accuracy favored GNN (0.670 versus 0.655). Median absolute error was lowest for MLP and highest for GNN, despite the GNN's lower mean absolute and squared errors, suggesting that its advantage was concentrated among larger errors. Mean signed error was small for all models and closest to zero for GNN.

On the held-out set, most correlation and error metrics improved relative to cross-validation. These comparisons reflect both refitting on the full development set and a different evaluation distribution, so the improvement cannot be attributed to training-set size alone. PR-AUC decreased for all models, and F_1 decreased for XGB and MLP. Predicted-positive rates fell from 16.6–30.0% in cross-validation to 5.6–14.7% on the held-out set, where approximately 82% of molecules were negative at the 0.7 threshold. The GNN nevertheless retained the highest held-out recall (0.53), MCC (0.505) and balanced accuracy (0.732).

E Comparison with DeepPSA

We compared DeepPSA's predicted probability of being "easy to synthesize" (Zhang et al., 2025) with the route-derived score and predictions from RetroTAC's Caruana ensemble. From the held-out set, we removed 51 molecules found in DeepPSA's training data and a further 378 used for ensemble selection, leaving 1,803 for evaluation. Pearson and Spearman correlations used continuous outputs.

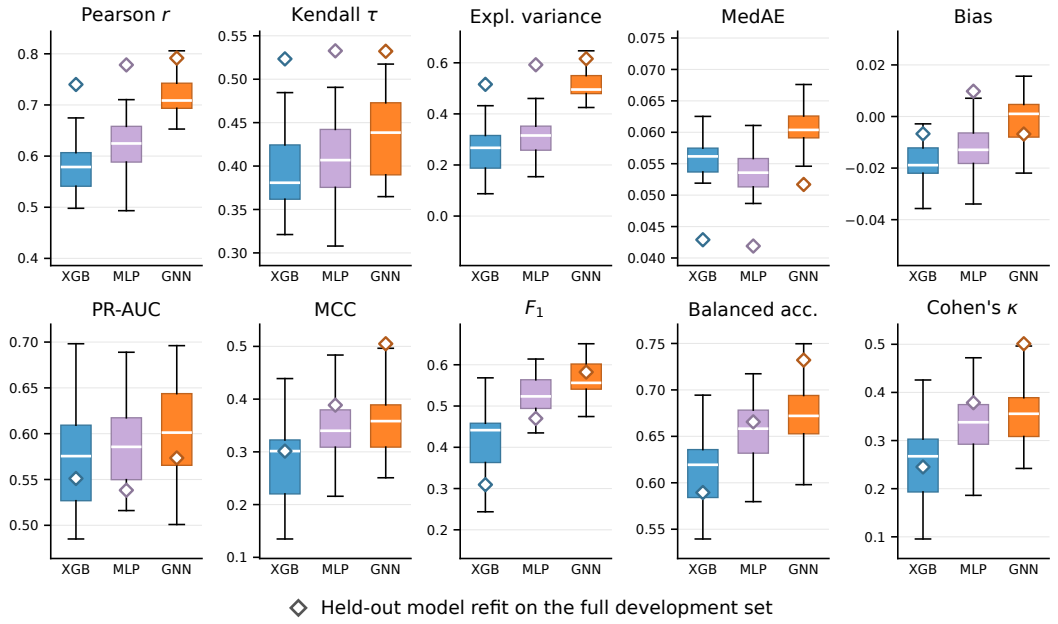


Figure 10: Additional performance metrics across cross-validation and held-out evaluation. Boxes summarize the 25 outer folds; open diamonds denote models refitted on the full development set and evaluated on 2,232 held-out molecules. The top row shows regression and ranking metrics; the bottom row shows classification metrics at a threshold of 0.7, with PR-AUC calculated from continuous predictions.

Table 12: Agreement with DeepPSA on the 1,803 held-out molecules remaining after exclusions for training overlap and ensemble selection. Pearson and Spearman correlations use continuous outputs; Kendall's τ uses outputs thresholded at 0.7 for the route-derived score and RETROTAC, and 0.5 for DeepPSA.

Comparison	Pearson (r)	Spearman (ρ)	Kendall (τ)
Route-derived score vs. DeepPSA	0.2392	0.3724	0.2539
Caruana prediction vs. DeepPSA	0.3144	0.4919	0.2485

For Kendall's τ, route scores and RETROTAC predictions were thresholded at 0.7 and DeepPSA probabilities at 0.5.

Continuous-output correlations with DeepPSA were higher for RETROTAC predictions ($r = 0.314$, $\rho = 0.492$) than for the route-derived scores ($r = 0.239$, $\rho = 0.372$; Table 12). The corresponding thresholded Kendall correlations were similar (0.249 and 0.254, respectively).

F ADDITIONAL EXAMPLE ROUTES

Figure 11 shows five routes generated by ShallowTree, with route-derived synthesizability scores ranging from 0.6 to 0.8. Target PROTACs are outlined in blue, intermediates in grey, and purchasable building blocks in green. The examples include sequential fragment assembly and convergent routes that join separately prepared intermediates, illustrating how different route architectures can achieve similar scores through their combinations of route depth and fragment balance.

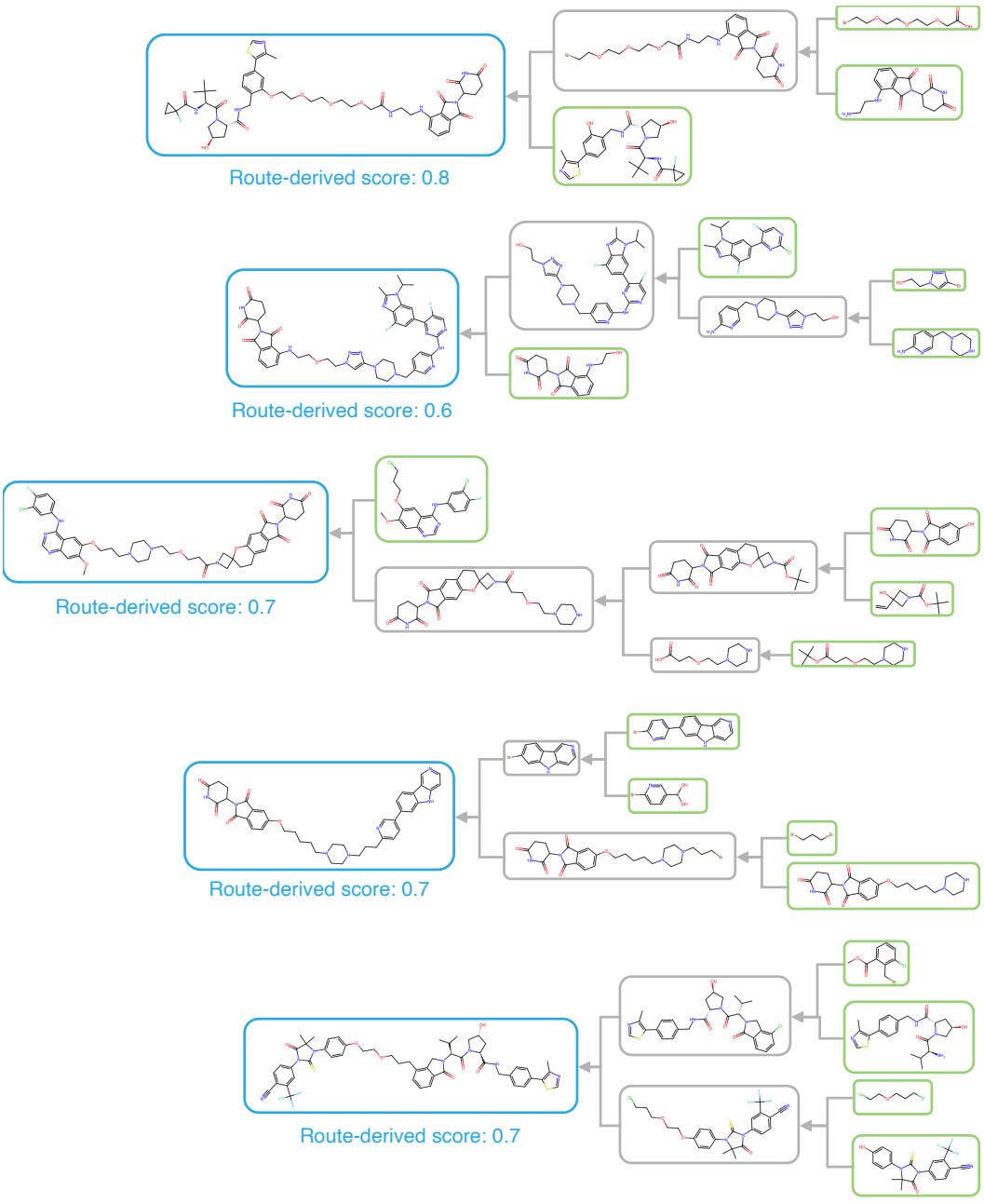


Figure 11: Five retrosynthesis routes generated by ShallowTree.