Original Article

DOI: 10.5582/bst.2026.01254

Selectivity-aware dual-target cholinesterase optimization using a novel model SPaR-GNN

Bocheng Zhang[1,§], Shaobo Bian[2,§], Yahui Liu[3], Kenji Karako[4,*]

[1] Graduate School of Frontier Sciences, Department of Integrated Biosciences, The University of Tokyo, Kashiwa, Japan;
[2] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, China;
[3] Graduate School of Life Sciences, Tohoku University, Sendai, Japan;
[4] Department of Surgery, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan.

SUMMARY: We developed SPaR-GNN (Selectivity Pair-aware Ranking Graph Neural Network), a multitask molecular learning framework for dual-target cholinesterase optimization. SPaR-GNN jointly predicts acetylcholinesterase (AChE) activity, butyrylcholinesterase (BuChE) activity, and molecular selectivity while incorporating explicit pair-aware supervision from structurally similar analogs and selectivity-cliff pairs. In addition, we integrated SPaR-GNN into a goal-driven molecular optimization workflow for candidate generation and prioritization. Experimental results showed that SPaR-GNN performed satisfactorily in selectivity-oriented prediction, and particularly in local analog ranking and direction-aware pair-sensitive benchmarks. The optimization framework further demonstrated that template-based candidate generation combined with SPaR-GNN scoring can be used to prioritize candidate molecules according to predefined dual-target optimization objectives. Together, these results demonstrate that SPaR-GNN provides an effective and traceable computational framework for dual-target cholinesterase inhibitor discovery and optimization in Alzheimer's disease.

Keywords: Alzheimer's disease, acetylcholinesterase, butyrylcholinesterase, graph neural network, multitask learning, drug discovery, molecular optimization

1. Introduction

Alzheimer's disease (AD), the most common form of dementia, is characterized by progressive memory loss and cognitive decline (*1,2*). Its pathological features include amyloid-β deposition, Tau hyperphosphorylation, neuroinflammation, neuronal loss, and synaptic dysfunction (*1,3,4*). Despite advances in understanding AD pathology, effective therapies remain limited. Approved cholinesterase inhibitors, including donepezil, galantamine, and rivastigmine, primarily enhance cholinergic neurotransmission rather than directly modifying the underlying disease pathology, and their therapeutic efficacy may decline with disease progression (*5*). Both acetylcholinesterase (AChE) and butyrylcholinesterase (BuChE) are involved in cholinergic dysfunction during AD progression (*6,7*), motivating the development of dual-target cholinesterase inhibitors.

Conventional medicinal chemistry approaches to dual-target compound discovery and optimization often involve extensive trial-and-error and experimental costs (*8*). Computational approaches that involve molecular representations and prioritize promising candidates therefore provide an attractive strategy to improve screening efficiency (*9*). However, dual-target cholinesterase modeling requires simultaneous consideration of AChE activity, BuChE activity, and target selectivity, making candidate prioritization sensitive to structural variations. Conventional fingerprint-based machine learning models and multilayer perceptrons (MLPs) rely on predefined molecular representations, whereas graph neural networks (GNNs) directly represent molecules as atom–bond graphs and can incorporate structure-dependent molecular features (*10-13*).

Beyond molecule-level prediction, lead optimization frequently requires differentiation of closely related analogs. This is particularly challenging for activity and selectivity cliffs, where minor structural modifications can produce disproportionately large changes in biological activity or target preference (*14*). Consequently, good global prediction performance does not necessarily ensure reliable prioritization of structurally similar compounds.

To address these challenges, we developed SPaR-GNN (Selectivity Pair-aware Ranking Graph Neural Network), a multitask molecular learning framework for dual-target cholinesterase modeling. SPaR-GNN jointly

models AChE activity, BuChE activity, and selectivity while incorporating pair-aware supervision to include local structure–selectivity relationships among related analogs. This combines multitask learning across related prediction tasks with explicit learning of pairwise relationships relevant to analog prioritization (*15,16*).

In this study, we curated and standardized an AChE/BuChE dataset and created pair-sensitive benchmarks consisting of structurally similar pairs, selectivity cliffs, and strict selectivity cliffs (AChE–BuChE polarity reversals). We developed SPaR-GNN to evaluate both molecule-level prediction and local analog discrimination, and we also integrated the model into a goal-driven molecular optimization workflow to prioritize AChE-oriented or BuChE-oriented candidates. The curated dataset, benchmark splits, model implementation, and optimization workflow are publicly available through our GitHub repository (*https://github.com/Eucliwood-bsb/SPaR-GNN*).

2. Materials and Methods

2.1. Dataset construction

2.1.1. Dataset collection, standardization, and quality control

Bioactivity data for AChE and BuChE were collected from ChEMBL (*17*) version 36 and BindingDB (*18*). Only quantitative half-maximal inhibitory concentration (IC_{50}) measurements were retained. Records lacking valid molecular structures, target information, numerical activity values, or activity units were excluded. Target names and organism annotations were standardized, activity values were converted to IC_{50} (nM) and transformed into pIC_{50}, and molecular structures were standardized using RDKit (*19*) to generate canonical SMILES representations. For molecule–target pairs with multiple measurements, median pIC_{50} values were used as representative activity labels. To improve label reliability, replicated measurements were further evaluated using the pIC_{50} range, and labels were retained when only a single measurement was available or when range$|pIC_{50} \leq 1.5$. The resulting dataset was organized in a multitask format containing AChE activity, BuChE activity, and target-availability masks. Although this procedure reduces variability among replicated measurements, systematic differences arising from assay protocols, enzyme preparations, substrate conditions, or other experimental settings across public databases cannot be completely eliminated.

Molecules with activity measurements for both targets were used to construct the dual-target subset. Selectivity was defined as

$$Selectivity = pIC50_{AChE} - pIC50_{BuChE}$$

Molecules were classified as AChE-selective (Selectivity ≥ 1.0), BuChE-selective (Selectivity ≤ -1.0), or balanced ($-1.0 <$ Selectivity < 1.0). Because selectivity is derived from the difference between two independently measured activities, residual assay-to-assay variability may propagate into the selectivity labels even after activity standardization and quality filtering.

2.1.2. Identification of structurally similar pairs and creation of selectivity-cliff benchmarks

Pairwise molecular relationships were identified among molecules with dual-target activity labels. Molecular similarity was calculated using Morgan fingerprints and Tanimoto similarity. Molecule pairs with a Tanimoto similarity greater than or equal to 0.60 were defined as structurally similar pairs. This procedure produced 24,452 structurally similar pairs.

For each structurally similar pair (i,j), the pairwise selectivity difference was defined as

$$\Delta y_{ij}^{S} = y_i^{S} - y_j^{S}. \quad (1)$$

Each pair was additionally annotated according to the selectivity classes of its two endpoints. The 24,452 structurally similar pairs consisted of 4,104 AChE-selective–AChE-selective pairs, 4,344 AChE-selective–balanced pairs, 705 AChE-selective–BuChE-selective pairs, 1,539 BuChE-selective–BuChE-selective pairs, 2,928 BuChE-selective–balanced pairs, and 10,832 balanced–balanced pairs.

Selectivity cliffs were defined as structurally similar pairs with endpoints that belonged to different selectivity classes and with an absolute selectivity difference greater than or equal to 1.0 log unit:

$$\left|\Delta y_{ij}^{S}\right| \geq 1.0. \quad (2)$$

The eligible cross-class combinations were AChE-selective–balanced, BuChE-selective–balanced, and AChE-selective–BuChE-selective. A total of 7,977 structurally similar pairs satisfied these endpoint-class and selectivity-difference criteria and constituted the selectivity-cliff benchmark. These included 4,344 AChE-selective–balanced pairs, 2,928 BuChE-selective–balanced pairs, and 705 AChE-selective–BuChE-selective pairs.

Strict selectivity cliffs were defined as the AChE-selective–BuChE-selective subset of the selectivity-cliff benchmark. These pairs represent reversals between the two opposing target-selectivity classes. A total of 705 pairs met this definition and constituted the strict selectivity-cliff subset. Because an AChE-selective endpoint has $y^{S} \geq 1.0$ and a BuChE-selective endpoint has $y^{S} \leq -1.0$, every strict selectivity cliff necessarily has an absolute selectivity difference of at least 2.0 log units.

The full set of structurally similar pairs was used as the basic pair set for pair-aware learning and similar-pair

evaluation. Pair-type-dependent weights were assigned hierarchically: structurally similar pairs that were not selectivity cliffs were given a weight of 1.0, selectivity cliffs were given a weight of 3.0, and strict selectivity cliffs were given a weight of 5.0. Strict-cliff weights replaced, rather than accumulated with, the lower pair-type weights.

Activity cliffs were annotated separately for AChE and BuChE. Structurally similar pairs exhibiting an activity difference greater than or equal to 1.0 log unit for the corresponding target were labeled as target-specific activity cliffs and used only in the relevant downstream analyses.

2.1.3. Dataset splitting strategy and pairwise leakage control

A pair-first scaffold-closure splitting strategy was used for the pair-sensitive experiments. First, 24,452 structurally similar molecular pairs were identified from the standardized dual-target dataset using a Morgan fingerprint Tanimoto similarity threshold of 0.60. Each pair was then annotated as an ordinary structurally similar pair, a selectivity cliff, or a strict selectivity cliff according to the definitions provided in Section 2.1.2. Pair identification and pair-type annotation were completed before assigning molecules to the training, validation, and test partitions.

To prevent the two endpoints of a structurally similar pair from being assigned to different partitions, molecules connected through the retained pair relationships were grouped into connected components. For the scaffold-closure split, Bemis–Murcko scaffold relationships were additionally incorporated into component construction. Molecules sharing the same Bemis–Murcko scaffold, as well as molecules connected through retained structurally similar pair relationships, were placed in the same closure component.

The resulting closure components were assigned exclusively to the training, validation, or test partition. Molecules inherited the partition assignment of their corresponding component, and each molecular pair was assigned to the same partition as its two endpoints. Only pairs with two endpoints in the same partition were retained in the corresponding pair subset.

In accordance with this protocol, the selectivity-cliff benchmark contained 5,863 training pairs, 1,060 validation pairs, and 1,054 test pairs. The strict selectivity-cliff subset contained 424 training pairs, 147 validation pairs, and 134 test pairs. The complete similar-pair test subset contained 3,373 pairs and was used to calculate similar-pair direction accuracy.

This procedure ensured that no retained molecular pair spanned different data partitions. In the scaffold-closure setting, it additionally ensured that no Bemis–Murcko scaffold was shared among the training, validation, and test sets, thereby providing a more stringent assessment of generalization to unseen chemotypes. This was used as the primary protocol for the pair-sensitive results.

2.2. Model: SPaR-GNN (Selectivity Pair-aware Ranking GNN)

As shown in Figure 1, we propose SPaR-GNN to predict the dual-target molecular properties of cholinesterase-related compounds (*20*). Let a molecule be represented as a graph $G = (V,E)$, where atoms are nodes and bonds are edges. For molecule i, the target values are denoted by y_i^A and y_i^B, corresponding to the pIC_{50} values measured against AChE and BuChE, respectively. For molecules with both target labels, we define the selectivity as $y_i^S = y_i^A - y_i^B$.

This formula reflects the degree to which a compound is biased towards AChE over BuChE. In practice, not every molecule has complete annotations for both targets. Therefore, we introduced the binary masks $m_i^A \in \{0,1\}$ and $m_i^B \in \{0,1\}$ to indicate whether the corresponding activity labels are observed. The selectivity mask is defined as $m_i^S = m_i^A \cdot m_i^B$ so that selectivity supervision is only applied to dual-labeled molecules.

The learning objective in our setting was not limited to accurate prediction of absolute activity. Instead, we sought to simultaneously model: (i) AChE activity, (ii) BuChE activity, and (iii) selectivity-related differences between the two targets. This multitask formulation is motivated by the fact that selective inhibitor design depends on both shared structure–activity patterns and target-specific deviations. Moreover, medicinal chemistry decisions often rely on relative comparisons among structurally similar analogs rather than isolated point estimates for individual molecules (*21*).

To better align training with this optimization-oriented setting, SPaR-GNN combines single-molecule regression with pair-level supervision derived from structurally similar molecular pairs (*22*), including selectivity cliff pairs. As summarized in the overall architecture in Figure 1, this design enables the model to learn not only global activity trends but also local structure–selectivity relationships that are especially important to analog prioritization.

2.3. Selectivity-aware multitask GINE

The molecular encoder of SPaR-GNN is based on a residual Graph Isomorphism Network with Edge features (GINE) (*23*). Atom and bond features are first projected into a shared hidden space and processed through stacked residual GINE layers (*24*). Representations from multiple encoder layers are subsequently integrated using an attention-like layer-fusion mechanism, followed by global mean and max pooling to obtain a shared molecular embedding. Detailed encoder formulations, including feature projection, residual message passing,

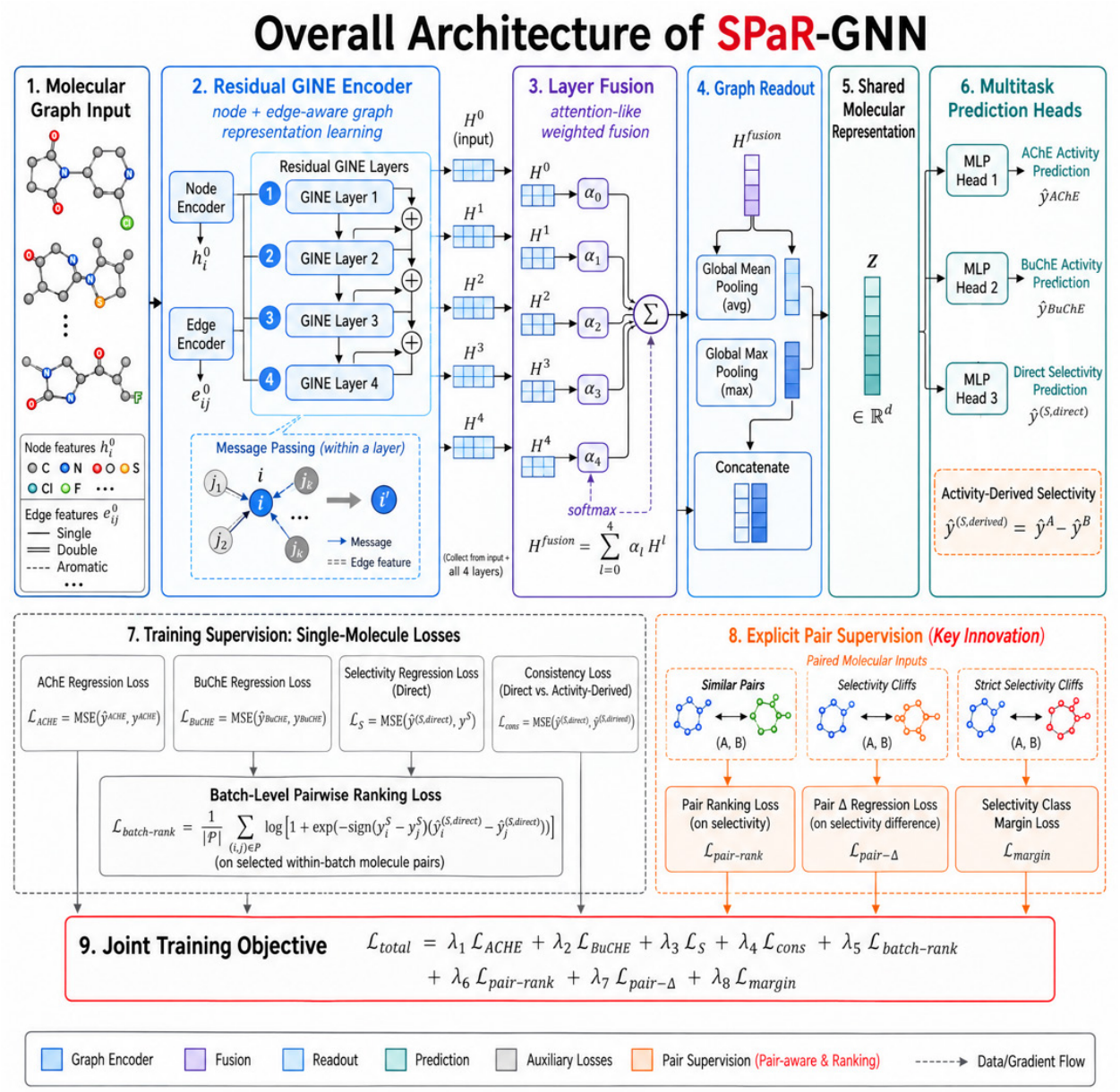


Figure 1. Overall architecture of SPaR-GNN.

layer fusion, and graph-level pooling, are provided in Supplementary Figure S1 (*https://www.biosciencetrends.com/action/getSupplementalData.php?ID=320*).

The resulting shared molecular representation is passed to three task-specific prediction heads for AChE activity, BuChE activity, and direct selectivity. SPaR-GNN therefore provides two distinct selectivity estimates. The direct selectivity estimate is generated by the dedicated selectivity head $\hat{y}^{S,\mathrm{direct}}$. In contrast, the activity-derived selectivity estimate is calculated deterministically from the two target-specific activity predictions $\hat{y}^{S,\mathrm{derived}}$. Thus, the activity-derived estimate is not generated from the direct selectivity output. The dedicated direct-selectivity branch explicitly learns the measured selectivity endpoint, whereas the activity-derived branch preserves the relationship between selectivity and the two target-specific activity predictions. The two estimates are connected during training through the selectivity-consistency objective described below. This dual-view approach is intended to facilitate selectivity-oriented prediction and downstream molecular optimization. Detailed structures of the prediction heads are provided in Supplementary Figure S2 (*https://www.biosciencetrends.com/action/getSupplementalData.php?ID=320*).

2.4. Explicit pair supervision and the training objective

The explicit pair supervision module of SPaR-GNN is shown in Figure 2. The training objective combines single-molecule multitask regression, selectivity consistency, batch-level ranking, and explicit pair supervision to jointly model absolute activity and local structure–selectivity relationships.

For individual molecules, masked mean squared error losses are applied to AChE, BuChE, and direct selectivity predictions:

$$L_A = \frac{\sum_i m_i^A \left(\hat{y}_i^A - y_i^A\right)^2}{\sum_i m_i^A},$$
$$L_B = \frac{\sum_i m_i^B \left(\hat{y}_i^B - y_i^B\right)^2}{\sum_i m_i^B}, \quad (3)$$
$$L_S = \frac{\sum_i m_i^S \left(\hat{y}_i^{S,direct} - y_i^S\right)^2}{\sum_i m_i^S}.$$

Consistency between direct selectivity prediction and the activity-derived estimate $\hat{y}_i^{S,derived}$ is enforced by

$$L_{cons} = \frac{\sum_i m_i^S \left(\hat{y}_i^{S,direct} - \hat{y}_i^{S,derived}\right)^2}{\sum_i m_i^S}. \quad (4)$$
$$\hat{y}_i^{S,derived} = \hat{y}_i^A - \hat{y}_i^B$$

A symmetric stop-gradient formulation is used in implementation to reduce the domination of either prediction branch.

For batch-level selectivity ranking (*25*), dual-labeled molecule pairs *(i,j)* are selected when

$$|y_i^S - y_j^S| \geq \gamma. \quad (5)$$

The corresponding ranking loss is

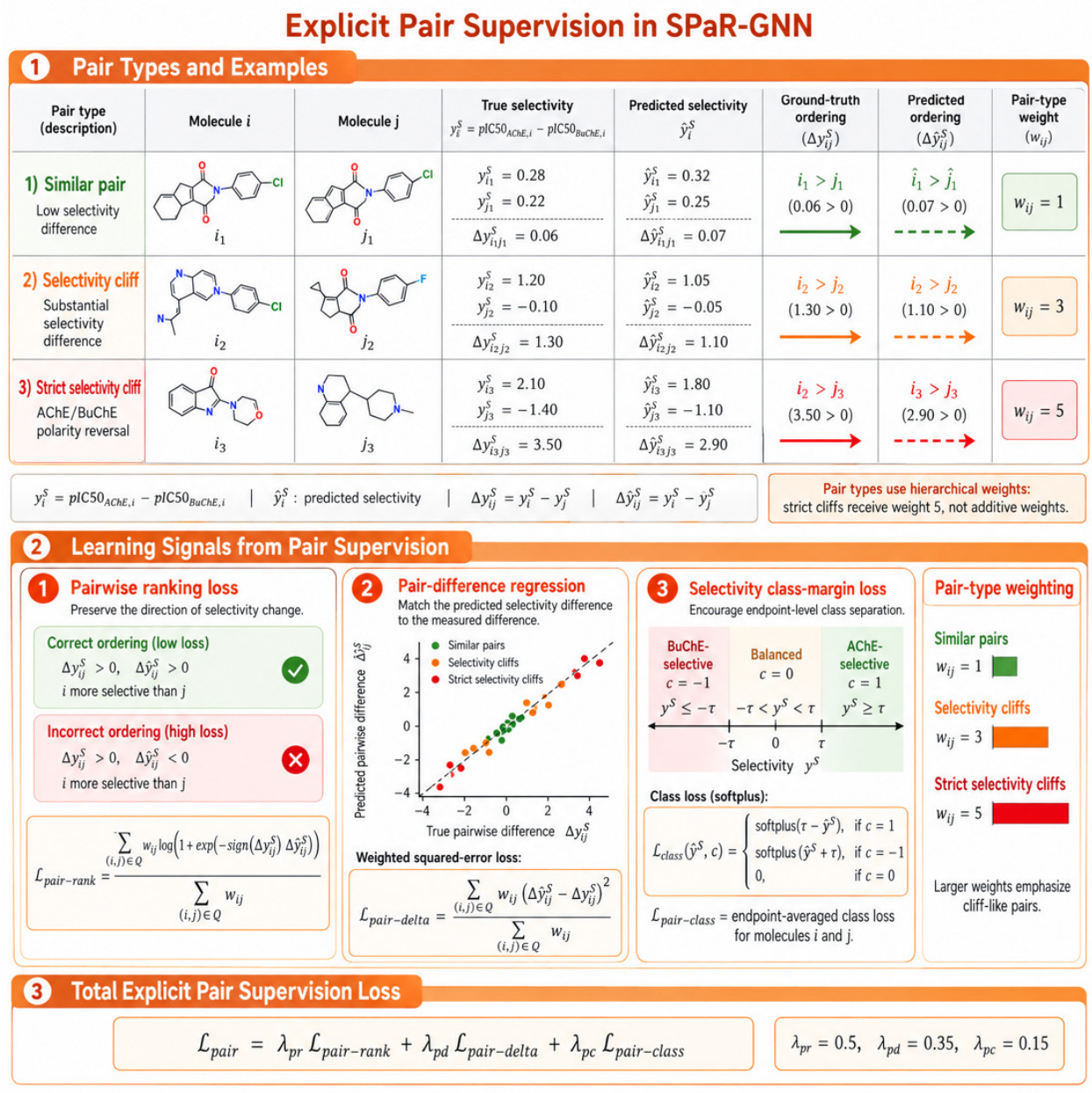


Figure 2. Explicit pair supervision in SPaR-GNN. Structurally similar pairs, selectivity cliffs, and strict selectivity cliffs are used for pairwise ranking, selectivity-difference regression, and endpoint-level class-margin supervision. Pair-type weights of 1.0, 3.0, and 5.0 are assigned hierarchically to ordinary structurally similar pairs, selectivity cliffs, and strict selectivity cliffs, respectively.

$$L_{\text{batch_rank}} = \frac{1}{|P|}\sum_{(i,j)\in P}\log\left(1+\exp\left[-\text{sign}(y_i^S-y_j^S)(\hat{y}_i^{S,direct}-\hat{y}_j^{S,direct})\right]\right), \quad (6)$$

where P denotes the selected within-batch pairs.

Explicit pair supervision is applied to the complete structurally similar pair set, with selectivity cliffs and strict selectivity cliffs treated as increasingly difficult nested subsets (*26*). Let Q denote the explicit training-pair set. Ordinary structurally similar pairs, selectivity cliffs, and strict selectivity cliffs are given hierarchical pair weights of 1.0, 3.0, and 5.0, respectively. For each pair *(i,j)*,

$$\Delta y_{ij}^S = y_i^S - y_j^S, \quad (7)$$

and

$$\Delta\hat{y}_{ij}^{S,direct} = \hat{y}_i^{S,direct} - \hat{y}_j^{S,direct}. \quad (8)$$

The pairwise ranking loss is defined as

$$L_{\text{pair_rank}} = \frac{\sum_{(i,j)\in Q} w_{ij}\log\left(1+\exp\left[-\text{sign}(\Delta y_{ij}^S)\Delta\hat{y}_{ij}^{S,direct}\right]\right)}{\sum_{(i,j)\in Q} w_{ij}}, \quad (9)$$

where Q denotes the explicit training-pair set and w_{ij} is the hierarchical pair-type weight. Specifically, $w_{ij}=1.0$ for structurally similar pairs that are not selectivity cliffs, $w_{ij}=3.0$ for selectivity cliffs that are not strict cliffs, and $w_{ij}=5.0$ for strict selectivity cliffs. The weights are mutually exclusive and are not accumulated across nested pair categories.

Pairwise selectivity differences are additionally optimized through

$$L_{\text{pair_delta}} = \frac{\sum_{(i,j)\in Q} w_{ij}\left(\Delta\hat{y}_{ij}^{S,direct}-\Delta y_{ij}^S\right)^2}{\sum_{(i,j)\in Q} w_{ij}}. \quad (10)$$

A class-margin loss is further used to distinguish AChE-selective, balanced, and BuChE-selective compounds according to threshold τ:

$$L_{\text{class}} = \begin{cases}\log\left(1+\exp(\tau-\hat{y}^{S,direct})\right), & c=1,\\ \log\left(1+\exp(\hat{y}^{S,direct}+\tau)\right), & c=-1,\\ 0, & c=0,\end{cases} \quad (11)$$

where $c\in\{-1,0,1\}$ denotes BuChE-selective, balanced, and AChE-selective classes, respectively.

The explicit pair supervision term is

$$L_{\text{pair}} = \lambda_{\text{pr}}L_{\text{pair_rank}}+\lambda_{\text{pd}}L_{\text{pair_delta}}+\lambda_{\text{pc}}L_{\text{pair_class}}, \quad (12)$$

Where $L_{\text{pair_class}}$ is the endpoint-averaged class-margin loss for each explicit pair.

Finally, the overall objective is

$$L_{\text{total}} = \lambda_A L_A+\lambda_B L_B+\lambda_S L_S+\lambda_C L_{\text{cons}}+\lambda_R L_{\text{batch_rank}}+L_{\text{pair}}, \quad (13)$$

Thus, SPaR-GNN jointly optimizes molecule-level activity and selectivity prediction, relative selectivity ordering, and pairwise discrimination of structurally related analogs.

2.5. Goal-driven molecular optimization framework

To extend SPaR-GNN from predictive modeling to analog prioritization, we developed a goal-driven molecular optimization framework that combines template-based local molecular editing with SPaR-GNN-based scoring (Figure 3) (*27*). Given an input molecule and a user-defined optimization goal G, the framework first generates structurally related candidates and applies prediction-independent structural and physicochemical filters. SPaR-GNN is then used to evaluate the retained candidates and the input molecule. The resulting predictions are used to apply prediction-dependent goal-specific constraints, after which feasible candidates are scored and ranked according to the specified objective.

Rather than performing unrestricted de novo generation, the framework searches a finite set of locally edited analogs $C(x_0)=\{x_1,x_2,\ldots,x_n\}$ while preserving structural proximity to the starting scaffold (*28*). The goal specification G defines optimization preferences and constraints, including target activity, selectivity, structural similarity, and physicochemical properties (*29*).

For each candidate x, SPaR-GNN provides predicted AChE activity $\hat{y}^A(x)$, BuChE activity $\hat{y}^B(x)$, derived selectivity $\hat{y}^{S,\text{derived}}(x)$, and direct selectivity $\hat{y}^{S,\text{direct}}(x)$. Structural similarity to the starting molecule is quantified using the Morgan fingerprint Tanimoto similarity, $\text{Sim}(x,x_0)$ (*30*).

Candidate optimization is formulated as a constrained ranking problem in which hard constraints determine candidate admissibility and soft preferences define goal-dependent prioritization. This formulation allows AChE-oriented, BuChE-oriented, or balanced dual-target objectives to be specified by changing the goal configuration without retraining the predictive model. Candidate generation, filtering, and ranking are described in the following section.

2.6. Template-based candidate generation, filtering, and ranking

To ensure chemical interpretability and controllability, candidate generation was performed through template-based local editing rather than unconstrained molecular generation (*31*). Given an input molecule x_0, a predefined library of medicinal chemistry-inspired transformation rules was applied to generate a structurally related candidate pool $C(x_0)$ (*32,33*). The transformations included commonly used local modifications such as

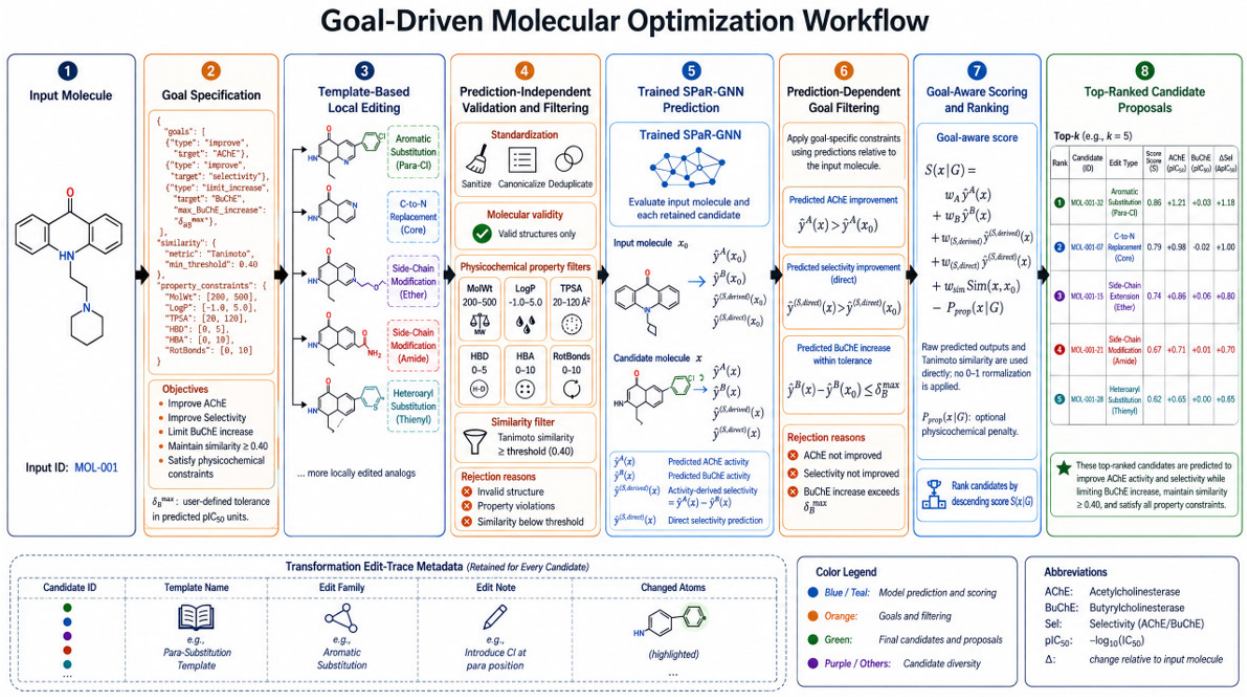


Figure 3. Goal-driven molecular optimization workflow based on SPaR-GNN. Starting from an input molecule and a user-defined objective, locally edited analogs are generated and screened using prediction-independent structural and physicochemical constraints. SPaR-GNN then evaluates the retained candidates and the input molecule. The resulting predictions are used for prediction-dependent goal-specific filtering, followed by direct weighted scoring and final candidate ranking.

aromatic substitution, bioisosteric replacement, and side chain modification. Each generated candidate was standardized, sanitized, deduplicated, and associated with an edit trace describing the corresponding structural modification. Detailed transformation rules and examples are provided in Supplementary Figure S3 (*https://www.biosciencetrends.com/action/getSupplementalData.php?ID=320*).

After generation, candidate molecules were standardized, sanitized, and deduplicated. Prediction-independent filtering was then applied using molecular validity, structural similarity, and physicochemical property constraints. These filters did not require SPaR-GNN outputs.

SPaR-GNN was subsequently applied to the candidates retained after prediction-independent filtering and to the input molecule. The model provided predicted AChE activity, predicted BuChE activity, derived selectivity, and direct selectivity. Goal-specific, prediction-dependent constraints were then evaluated relative to the corresponding predictions for the input molecule. Depending on the optimization objective, candidates could be required to improve predicted AChE activity, improve predicted selectivity, limit an undesirable increase in predicted BuChE activity, or satisfy the corresponding BuChE-oriented criteria (Supplementary Figure S4 (*https://www.biosciencetrends.com/action/getSupplementalData.php?ID=320*).

Candidates satisfying both the prediction-independent and prediction-dependent constraints were considered feasible. A goal-dependent score was then calculated to integrate the model predictions with structural similarity and optional physicochemical penalties:

$$S(x|G) = w_A\hat{y}^A(x) + w_B\hat{y}^B(x) + w_{S,\text{derived}}\hat{y}^{S,\text{derived}}(x) + w_{S,\text{direct}}\hat{y}^{S,\text{direct}}(x) + w_{\text{sim}}\text{Sim}(x,x_0) - P_{\text{prop}}(x|G). \quad (14)$$

The user-defined optimization objective is denoted, and the relative contributions of the corresponding prediction and similarity terms are controlled. Here, $P_{\text{prop}}(x|G)$ represents optional penalties for deviations from preferred physicochemical ranges.

The feasible candidates were ranked in descending order according to $S(x|G)$, and the top-ranked molecules were retained as optimization proposals. Because the score weights and filtering criteria can be adjusted according to the desired target preference, the same candidate-generation framework can support AChE-oriented, BuChE-oriented, or balanced dual-target optimization without retraining the predictive model. Detailed goal configurations, physicochemical thresholds, candidate-level scores, and edit-trace information are provided in the Supplementary Materials (*https://www.biosciencetrends.com/action/getSupplementalData.php?ID=320*).

2.7. Implementation details and hyperparameter settings

All experiments were implemented in Python using

PyTorch and PyTorch Geometric and trained on an NVIDIA RTX 5070 GPU. To assess robustness to random initialization and stochastic training, key neural-model experiments were repeated using multiple independent random seeds. Key implementation details and hyperparameter settings are summarized in Supplementary Table S1 (*https://www.biosciencetrends.com/action/getSupplementalData.php?ID=320*).

Each molecule was represented as a graph with atoms as nodes and bonds as edges. Atom features included atomic number, hybridization, degree, formal charge, hydrogen count, aromaticity, and ring membership, while bond features included bond type, conjugation, ring membership, and stereochemistry. Categorical descriptors were one-hot encoded with an additional "unknown" category when necessary.

The SPaR-GNN encoder used four residual GINE layers with a hidden dimension of 128 and a dropout rate of 0.2. Node and edge features were projected into a shared hidden space, processed through stacked residual GINE blocks, fused using the attention-like layer fusion module, and aggregated by concatenated global mean and max pooling before being passed to three task-specific prediction heads.

Training used the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-5}. Molecule and pair batch sizes were 64 and 128, respectively. Training was performed for up to 300 epochs with early stopping patience of 35 epochs.

The loss weights were set to $\lambda_A = 1.0$, $\lambda_B = 1.0$, $\lambda_S = 0.8$, $\lambda_C = 0.5$, and $\lambda_R = 0.5$. For explicit pair supervision, $\lambda_{pr} = 0.5$, $\lambda_{pd} = 0.35$, and $\lambda_{pc} = 0.15$. Batch-level ranking used a margin threshold of $\gamma=0.5$ with up to 512 informative within-batch pairs. The selectivity class threshold was $\tau = 1.0$. Structurally similar pairs were defined using a Tanimoto similarity threshold of 0.60. Pair-type weights were assigned hierarchically: ordinary structurally similar pairs received a weight of 1.0, selectivity cliffs received a weight of 3.0, and strict selectivity cliffs received a weight of 5.0.

To ensure consistent validation-based checkpoint selection across neural models, a common selectivity-oriented validation framework was used for MLP, GNN, GINE, and SPaR-GNN. For models without a dedicated direct-selectivity head, the unavailable direct-selectivity term was omitted and the remaining terms were re-normalized. At each epoch, a mixed validation score was calculated based on direct selectivity Spearman correlation, activity-derived selectivity Spearman correlation, selectivity-cliff Spearman correlation, and strict selectivity-cliff direction accuracy:

$$S_{\mathrm{val}} = \frac{0.35\rho_{\mathrm{sel}}^{\mathrm{direct}} + 0.25\rho_{\mathrm{sel}}^{\mathrm{activity}} + 0.25\rho_{\mathrm{cliff}} + 0.15Acc_{\mathrm{dir}}^{\mathrm{strict}}}{w_{\mathrm{used}}}.$$

Here, w_{used} denotes the sum of weights corresponding to non-missing validation terms. For models without a direct selectivity prediction head, the corresponding term was omitted and the remaining terms were re-normalized. The checkpoint with the highest S_{val} was selected for final test-set evaluation.

For non-neural baselines such as XGBoost or random forest, no epoch-wise checkpoint selection was used. These models were trained with fixed hyperparameter settings and evaluated using the same validation and test partitions. Unless otherwise stated, all hyperparameters were fixed for the scaffold-closure experiments.

The selectivity-related metrics reported in tables were calculated using the direct selectivity prediction $\hat{y}^{S,\mathrm{direct}}$ in SPaR-GNN, and those metrics were calculated as $\hat{y}^{S,\mathrm{derived}}$ in the other baseline models without a dedicated selectivity head based on the predicted target activity.

3. Results

3.1. Dataset construction and characteristics

The resulting dual-target dataset contained 3,085 compounds, including 864 AChE-selective compounds (28.0%), 1,783 balanced compounds (57.8%), and 438 BuChE-selective compounds (14.2%). At the molecule level, 2,289, 397, and 399 compounds were assigned to the training, validation, and test partitions, respectively.

For molecules with activity measurements against both targets, selectivity was defined as:

$$y_i^S = pIC50_{\mathrm{AChE},i} - pIC50_{\mathrm{BuChE},i}.$$

To characterize local selectivity relationships, we enumerated all molecular pairs with a Morgan fingerprint Tanimoto similarity greater than or equal to 0.60. This procedure produced 24,452 structurally similar pairs. Each pair was annotated according to the selectivity classes of its two endpoints and its pairwise selectivity difference. The complete similar-pair set contained 4,104 AChE-selective–AChE-selective pairs, 4,344 AChE-selective–balanced pairs, 705 AChE-selective–BuChE-selective pairs, 1,539 BuChE-selective–BuChE-selective pairs, 2,928 BuChE-selective–balanced pairs, and 10,832 balanced–balanced pairs.

Selectivity cliffs were defined as structurally similar pairs with endpoints in different selectivity classes and with an absolute selectivity difference greater than or equal to 1.0 log unit. The eligible endpoint-class combinations were AChE-selective–balanced, BuChE-selective–balanced, and AChE-selective–BuChE-selective. A total of 7,977 pairs satisfied these criteria and constituted the selectivity-cliff benchmark. These included 4,344 AChE-selective–balanced pairs, 2,928 BuChE-selective–balanced pairs, and 705 AChE-selective–BuChE-selective pairs.

The 705 AChE-selective–BuChE-selective pairs were defined as strict selectivity cliffs because they represented polarity reversals between the two opposing

selectivity classes. These pairs form a nested subset of the 7,977 selectivity cliffs. No additional Tanimoto threshold beyond the general structurally similar-pair threshold of 0.60 was imposed for strict selectivity cliffs.

Under the final pair-first scaffold-closure protocol, the structurally similar pair relationships, together with Bemis–Murcko scaffold relationships, were used to form closure components before train/validation/test assignment. A total of 101 closure components were obtained and assigned exclusively to the training, validation, or test partition. Molecules inherited the split assignment of their corresponding component, and pairs were assigned to the same partition as their endpoints.

The selectivity-cliff benchmark contained 5,863 training pairs, 1,060 validation pairs, and 1,054 test pairs. The strict selectivity-cliff subset contained 424 training pairs, 147 validation pairs, and 134 test pairs. After component-level split assignment, all structurally similar pairs were filtered according to endpoint membership, yielding 3,373 structurally similar pairs with both endpoints in the test partition.

Accordingly, the three groups of pair-sensitive metrics were evaluated for distinct but nested test subsets. Similar-pair direction accuracy was calculated for all 3,373 structurally similar test pairs. Selectivity-cliff RMSE, Spearman correlation, and direction accuracy were calculated for all 1,054 test selectivity cliffs. Strict-cliff RMSE, Spearman correlation, and direction accuracy were calculated for the 134 test AChE-selective–BuChE-selective polarity-reversal pairs (Figure 4).

Table 1 summarizes the definitions, sizes, and uses of the pair subsets. The 7,977 selectivity cliffs can be divided into two mutually exclusive endpoint-class groups: 7,272 balanced-involving selectivity cliffs and 705 strict AChE-selective–BuChE-selective polarity-reversal pairs. The balanced-involving subset consists of 4,344 AChE-selective–balanced pairs and 2,928 BuChE-selective–balanced pairs. Thus, strict selectivity cliffs form a nested subset of the selectivity-cliff benchmark, whereas the three endpoint-class combinations form a mutually exclusive partition of all 7,977 selectivity cliffs.

Table 2 shows the endpoint-class distribution of the 7,977 selectivity cliffs after pair-first scaffold-closure splitting. AChE-selective–balanced cliffs were the most frequent, accounting for 4,344 pairs (54.46%), followed by BuChE-selective–balanced cliffs with 2,928 pairs (36.71%) and strict AChE-selective–BuChE-selective polarity reversals with 705 pairs (8.84%).

Because the structurally similar pair relationships

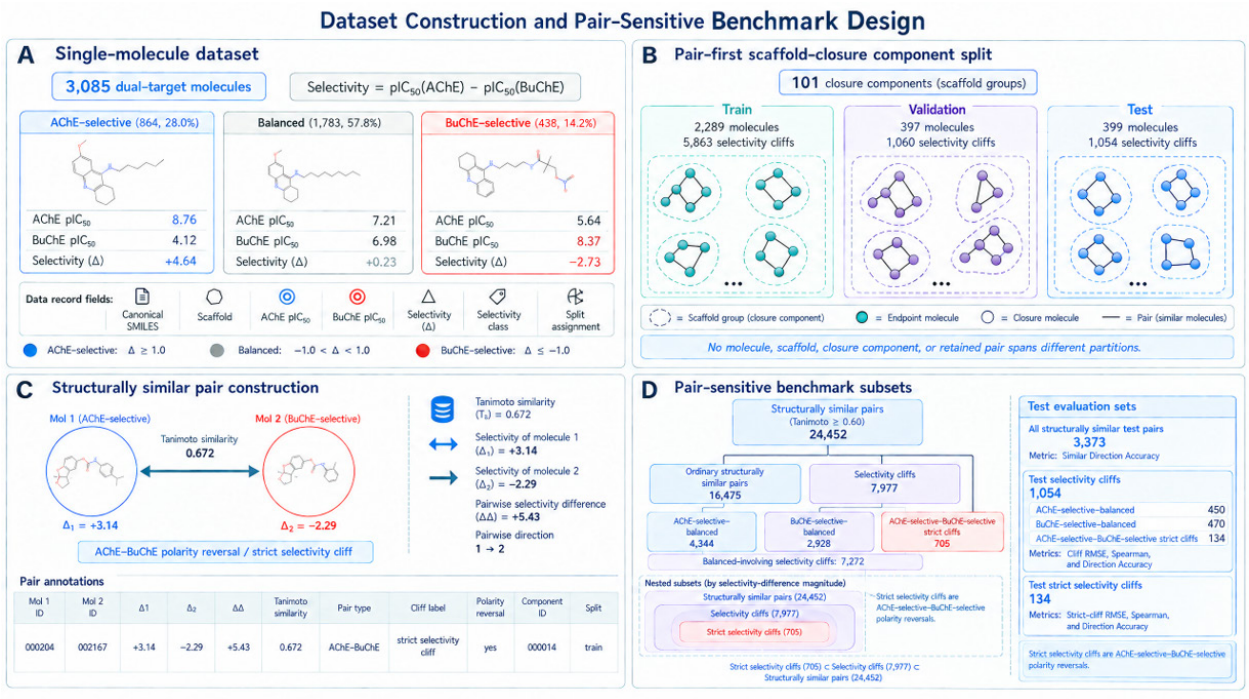


Figure 4. Overview of the cholinesterase selectivity dataset and pair-sensitive benchmark. A, The single-molecule dataset contains 3,085 compounds with dual-target AChE and BuChE potency labels. Molecular selectivity is defined as the difference between AChE and BuChE potency and is used to assign AChE-selective, balanced, or BuChE-selective classes. **B,** Pair-first scaffold-closure splitting based on 101 closure components. The selectivity-cliff benchmark contains 5,863 training, 1,060 validation, and 1,054 test pairs while preventing scaffold-, component-, and pair-endpoint-level leakage. **C,** Structurally similar pairs are identified using a Morgan fingerprint Tanimoto similarity threshold of 0.60 and are annotated with endpoint selectivity classes, pairwise selectivity differences, selectivity-cliff labels, and strict polarity-reversal labels. **D,** Relationships among the pair-sensitive subsets. The complete dataset contains 24,452 structurally similar pairs. Of these, 7,977 cross-class pairs with an absolute selectivity difference greater than or equal to 1.0 log unit are defined as selectivity cliffs. The 705 AChE-selective–BuChE-selective polarity-reversal pairs form the strict selectivity-cliff subset. The test subsets contain 3,373 structurally similar pairs, 1,054 selectivity cliffs, and 134 strict selectivity cliffs.

were identified before split assignment and subsequently incorporated into scaffold-closure component construction, we further assessed potential information leakage across the resulting data partitions (Table 3). With a scaffold-closure split, no train–test overlap was detected at the molecule, canonical or standardized SMILES, Bemis–Murcko scaffold, or pair-endpoint level. Pair assignments were fully consistent with the molecule-level split, and no endpoints of test pairs appeared among the training molecules, confirming the absence of pair-level leakage. This split therefore provided a more stringent assessment of generalization to unseen chemotypes.

3.2. SPaR-GNN improves selectivity prediction

To evaluate the predictive capability of SPaR-GNN, we compared SPaR-GNN against four representative baselines on the standard single-molecule multitask prediction task. The benchmark included a tree-based model using molecular fingerprints (XGB) (*34*), a fingerprint-based neural baseline (MLP) (*35*), a message-passing GNN (*12*), and a GINE (*36*) model with edge-aware graph convolutions. All models were evaluated using the same benchmark dataset with a scaffold split. Following the multitask setting described earlier, predictive performance in terms of AChE activity, BuChE activity, and selectivity are described below.

Table 4 summarizes the single-molecule multitask prediction results with a scaffold-closure split. As expected, prediction performance varied substantially across models when evaluated using unseen chemotypes. In the individual target tasks, XGB resulted in the best AChE performance, with the lowest RMSE (1.138 ± 0.019) and the highest Spearman correlation (0.723 ± 0.018). XGB also resulted in the lowest BuChE RMSE (1.149 ± 0.023), whereas MLP resulted in the highest BuChE Spearman correlation (0.642 ± 0.022). In contrast, SPaR-GNN did not outperform the baseline models in single-target activity prediction, indicating that its primary advantage does not lie in optimizing isolated AChE or BuChE regression accuracy.

A different pattern emerged for selectivity prediction. SPaR-GNN resulted in both the lowest selectivity RMSE

Table 1. Definitions, counts, and uses in evaluation for the molecular-pair subsets. Selectivity cliffs and strict selectivity cliffs are nested subsets of the structurally similar pair set

Pair subset	Definition	Total	Test	Use in the study
Structurally similar pairs	Morgan fingerprint Tanimoto similarity ≥ 0.60	24,452	3,373	Basic pair-aware learning set and Similar Dir. Acc. evaluation
Ordinary similar pairs	Structurally similar pairs that do not meet the selectivity-cliff definition	16,475	2,319	Explicit pair supervision with a pair weight of 1.0
Selectivity cliffs	Structurally similar cross-class pairs with $\lvert\Delta y^{S}_{ij}\rvert \geq 1.0$	7,977	1,054	Explicit pair supervision with a weight of 3.0 and cliff RMSE, Spearman, and direction accuracy
Balanced-involving selectivity cliffs	AChE-selective—balanced or BuChE-selective—balanced selectivity cliffs	7,272	920	Endpoint-class characterization of the cliff benchmark
Strict selectivity cliffs	AChE-selective—BuChE-selective polarity-reversal pairs within the selectivity-cliff set	705	134	Explicit pair supervision with a weight of 5.0 and strict-cliff RMSE, Spearman, and direction accuracy

Table 2. Distribution of the 7,977 selectivity cliffs after pair-first scaffold-closure splitting. Selectivity cliffs were defined as structurally similar cross-class pairs with an absolute selectivity difference greater than or equal to 1.0 log unit. Ratios within each split were calculated relative to the number of selectivity cliffs in the corresponding partition, whereas ratios of all pairs were calculated relative to all 7,977 selectivity cliffs

Split	Selectivity-cliff type	Count	Ratio within split	Ratio of all pairs
All	AChE-selective—balanced	4,344	54.46%	54.46%
All	BuChE-selective—balanced	2,928	36.71%	36.71%
All	AChE-selective—BuChE-selective polarity reversal	705	8.84%	8.84%
Train	AChE-selective—balanced	3,298	56.25%	41.34%
Train	BuChE-selective—balanced	2,141	36.52%	26.84%
Train	AChE-selective—BuChE-selective polarity reversal	424	7.23%	5.32%
Validation	AChE-selective—balanced	596	56.23%	7.47%
Validation	BuChE-selective—balanced	317	29.91%	3.97%
Validation	AChE-selective—BuChE-selective polarity reversal	147	13.87%	1.84%
Test	BuChE-selective—balanced	470	44.59%	5.89%
Test	AChE-selective—balanced	450	42.69%	5.64%
Test	AChE-selective—BuChE-selective polarity reversal	134	12.71%	1.68%

(1.071 ± 0.044) and the highest selectivity Spearman correlation (0.442 ± 0.054), outperforming all baseline models with the scaffold-closure split. MLP resulted in the next-best selectivity RMSE (1.145 ± 0.031), while XGB resulted in the second-highest selectivity Spearman correlation (0.366 ± 0.042). These results indicate that SPaR-GNN better preserves the relative relationship between AChE and BuChE activities when generalizing to unseen chemotypes, despite its weaker performance on the individual target regression tasks.

This distinction is important because accurate prediction of AChE or BuChE activity alone does not necessarily translate into accurate estimation of their relative selectivity. By incorporating selectivity-aware and pair-aware objectives during training, SPaR-GNN is designed to prioritize dual-target relationships rather than solely minimizing independent single-target prediction errors. The scaffold-split results therefore support the intended role of SPaR-GNN as a selectivity-oriented model, with its clearest advantage observed in selectivity prediction rather than individual target activity prediction.

Overall, these results demonstrate that the primary strength of SPaR-GNN lies in modeling selectivity across unseen chemotypes. This property is particularly relevant to downstream lead optimization, where reliable ranking and prioritization of compounds according to their selectivity profiles may be more informative than marginal improvements in single-target prediction accuracy.

3.3. SPaR-GNN captures pairwise selectivity relationships

Single-molecule metrics do not fully assess a model's ability to distinguish structurally related analogs with different selectivity profiles. We therefore evaluated all models on three nested pair-sensitive test subsets: structurally similar pairs, selectivity cliffs, and strict selectivity cliffs. Structurally similar pairs were defined using a Morgan fingerprint Tanimoto similarity threshold of 0.60. Selectivity cliffs were structurally similar cross-class pairs with an absolute selectivity difference greater than or equal to 1.0 log unit, whereas strict selectivity cliffs were the AChE-selective–BuChE-selective polarity-reversal subset.

The Similar Dir. Acc. metric was evaluated for all 3,373 structurally similar test pairs. The cliff RMSE, Spearman correlation, and direction accuracy metrics were evaluated for the 1,054 test selectivity cliffs. The strict-cliff RMSE, Spearman correlation, and direction accuracy metrics were evaluated for the 134 test strict selectivity cliffs.

To assess robustness to random initialization and stochastic training, all models were evaluated using multiple independent random seeds. The resulting mean ± standard deviation values are summarized in Table 5. With a scaffold-closure split, SPaR-GNN achieved the best mean performance across all seven pair-sensitive metrics. With the 3,373 structurally similar test pairs, SPaR-GNN achieved the highest direction accuracy (0.708 ± 0.021). With the 1,054 selectivity cliffs, it resulted in the lowest RMSE (2.083 ± 0.076), the highest Spearman correlation (0.646 ± 0.053), and the highest direction accuracy (0.742 ± 0.023). The same advantage was retained with the 134

Table 3. Leakage checks for the pair-first scaffold-closure split. All checks passed, indicating no molecule-, SMILES-, scaffold-, or pair-endpoint-level leakage between the training and test partitions

Level	Check	Key	Total/ Train	Test	Leakage/ Inconsistent	Status
Molecule	Cross-split uniqueness	mol_id	3085	-	0	PASS
Molecule	Train—test intersection	mol_id	2289	399	0	PASS
Molecule	Cross-split uniqueness	smiles_can	3085	-	0	PASS
Molecule	Train—test intersection	smiles_can	2289	399	0	PASS
Molecule	Cross-split uniqueness	smiles_std	3085	-	0	PASS
Molecule	Train—test intersection	smiles_std	2289	399	0	PASS
Scaffold	Cross-split uniqueness	scaffold	1318	-	0	PASS
Scaffold	Train—test intersection	scaffold	1077	138	0	PASS
Pair	Similar-pair split consistency with molecule split	-	24452	-	0	PASS
Pair endpoint	Test similar-pair endpoints in training molecules	mol_id	-	3373 pairs/397 endpoints	0	PASS
Selectivity cliff	Cliff split consistency with molecule split	-	7977	-	0	PASS
Cliff endpoint	Test-cliff endpoints in training molecules	mol_id	-	1054 pairs/230 endpoints	0	PASS

Table 4. Single-molecule multitask prediction results with a scaffold split. Results are expressed as the mean ± standard deviation on the test set over ten independent seeds. A lower RMSE indicates better regression accuracy, while a higher Spearman correlation indicates better rank consistency. The best result in each column is shown in bold

Model	AChE RMSE	AChE Spearman	BuChE RMSE	BuChE Spearman	Selectivity RMSE	Selectivity Spearman
XGB	**1.138 ± 0.019**	**0.723 ± 0.018**	**1.149 ± 0.023**	0.619 ± 0.029	1.152 ± 0.029	0.366 ± 0.042
MLP	1.208 ± 0.061	0.646 ± 0.025	1.180 ± 0.039	**0.642 ± 0.022**	1.145 ± 0.031	0.345 ± 0.044
GNN	1.471 ± 0.090	0.416 ± 0.067	1.231 ± 0.084	0.613 ± 0.051	1.247 ± 0.088	0.206 ± 0.115
GINE	1.488 ± 0.099	0.397 ± 0.064	1.278 ± 0.091	0.594 ± 0.051	1.214 ± 0.094	0.293 ± 0.089
SPaR-GNN	1.889 ± 0.221	0.523 ± 0.073	1.851 ± 0.225	0.590 ± 0.060	**1.071 ± 0.044**	**0.442 ± 0.054**

Table 5. Multi-seed comparison of baseline models and SPaR-GNN for pair-sensitive benchmarks with a scaffold-closure split. Similar direction accuracy was evaluated for all 3,373 structurally similar test pairs with a Morgan fingerprint Tanimoto similarity greater than or equal to 0.60. Cliff RMSE, Spearman correlation, and direction accuracy were evaluated for the 1,054 test selectivity cliffs. Strict-cliff metrics were evaluated for the 134 AChE-selective–BuChE-selective polarity-reversal pairs. Results are expressed as the mean ± standard deviation over independent random seeds. A lower RMSE indicates better pairwise regression accuracy, whereas a higher Spearman correlation and direction accuracy indicate better preservation of pairwise selectivity ordering. The best result in each metric column is shown in bold

Model	Similar Dir. Acc.	Cliff RMSE	Cliff Spearman	Cliff Dir. Acc.	Strict Cliff RMSE	Strict Cliff Spearman	Strict Cliff Dir. Acc.
XGB	0.618 ± 0.044	2.319 ± 0.051	0.380 ± 0.088	0.636 ± 0.053	3.063 ± 0.077	0.340 ± 0.076	0.775 ± 0.064
MLP	0.644 ± 0.029	2.282 ± 0.043	0.457 ± 0.061	0.659 ± 0.036	3.021 ± 0.075	0.413 ± 0.068	0.824 ± 0.044
GNN	0.560 ± 0.072	2.435 ± 0.091	0.208 ± 0.195	0.565 ± 0.085	3.409 ± 0.147	0.069 ± 0.201	0.609 ± 0.107
GINE	0.569 ± 0.068	2.408 ± 0.088	0.233 ± 0.167	0.573 ± 0.080	3.343 ± 0.147	0.136 ± 0.185	0.637 ± 0.107
SPaR-GNN	**0.708 ± 0.021**	**2.083 ± 0.076**	**0.646 ± 0.053**	**0.742 ± 0.023**	**2.813 ± 0.115**	**0.592 ± 0.076**	**0.880 ± 0.021**

Table 6. Ablation and baseline comparison of pair-sensitive benchmarks with a scaffold-closure split. All results are expressed as the mean ± standard deviation over 100 random seeds. Similar direction accuracy was evaluated for all 3,373 structurally similar test pairs. Cliff metrics were evaluated for the 1,054 test selectivity cliffs, and strict-cliff metrics were evaluated for the 134 test AChE-selective–BuChE-selective polarity-reversal pairs. Independent single-task GINE trains separate AChE and BuChE predictors and computes selectivity from their prediction difference. Parameter-matched GINE uses the same encoder and prediction heads as SPaR-GNN but removes pair-aware losses. Pairwise Morgan-XGB directly predicts pairwise selectivity differences from Morgan fingerprint pair features. A lower RMSE indicates better pairwise selectivity-difference regression, whereas a higher Spearman correlation and direction accuracy indicate better preservation of pairwise selectivity ordering. The best result in each metric column is shown in bold

Model/ Variant	Similar Dir. Acc.	Cliff RMSE	Cliff Spearman	Cliff Dir. Acc.	Strict Cliff RMSE	Strict Cliff Spearman	Strict Cliff Dir. Acc.
Full model SPaR-GNN	**0.708 ± 0.021**	**2.083 ± 0.076**	**0.646 ± 0.053**	**0.742 ± 0.023**	**2.813 ± 0.115**	**0.592 ± 0.076**	**0.880 ± 0.021**
Reviewer-requested baselines							
Independent single-task GINE	0.615 ± 0.048	2.368 ± 0.092	0.328 ± 0.163	0.630 ± 0.057	3.284 ± 0.134	0.216 ± 0.183	0.699 ± 0.105
Parameter-matched GINE w/o pair-aware loss	0.636 ± 0.054	2.307 ± 0.093	0.419 ± 0.156	0.658 ± 0.067	3.150 ± 0.152	0.306 ± 0.170	0.744 ± 0.107
Pairwise Morgan-XGB	0.596 ± 0.020	2.579 ± 0.026	0.053 ± 0.020	0.623 ± 0.024	3.443 ± 0.039	-0.095 ± 0.034	0.710 ± 0.039
SPaR-GNN component ablations							
w/o direct selectivity supervision	0.604 ± 0.059	2.286 ± 0.141	0.366 ± 0.175	0.617 ± 0.071	3.126 ± 0.212	0.283 ± 0.215	0.698 ± 0.117
w/o selectivity consistency loss	0.658 ± 0.047	2.196 ± 0.120	0.517 ± 0.139	0.684 ± 0.056	2.996 ± 0.171	0.451 ± 0.144	0.795 ± 0.073
w/o batch-level ranking loss	0.661 ± 0.045	2.195 ± 0.114	0.522 ± 0.139	0.686 ± 0.055	2.992 ± 0.167	0.470 ± 0.152	0.803 ± 0.074
w/o explicit pair ranking loss	0.661 ± 0.042	2.192 ± 0.108	0.524 ± 0.125	0.687 ± 0.049	2.981 ± 0.162	0.467 ± 0.152	0.798 ± 0.075
w/o pair-difference regression	0.631 ± 0.050	2.287 ± 0.093	0.409 ± 0.137	0.652 ± 0.060	3.083 ± 0.157	0.321 ± 0.153	0.751 ± 0.094
w/o cliff-pair weighting	0.649 ± 0.044	2.242 ± 0.098	0.478 ± 0.131	0.672 ± 0.054	3.058 ± 0.143	0.400 ± 0.143	0.774 ± 0.076

strict selectivity cliffs, as indicated by the lowest RMSE (2.813 ± 0.115), the highest Spearman correlation (0.592 ± 0.076), and the highest direction accuracy (0.880 ± 0.021).

These results show that SPaR-GNN improves both the numerical prediction and ordering of pairwise selectivity changes under scaffold-controlled generalization. In contrast to the conventional baselines, which exhibited larger performance losses with cliff-like pairs, SPaR-GNN maintained a stronger rank consistency and direction discrimination while also reducing pairwise regression error. The consistent advantage across RMSE, Spearman correlation, and direction accuracy indicates that the improvement is not restricted to a single aspect of pairwise prediction.

Notably, the fingerprint-based MLP was the strongest conventional baseline for most scaffold-split pair-sensitive metrics, outperforming the conventional GNN and GINE models. Nevertheless, SPaR-GNN consistently surpassed MLP across all seven metrics, including the more challenging strict selectivity-cliff benchmark. Together with the single-molecule results, these findings indicate that graph representation alone does not account for the performance of SPaR-GNN and they prompt further evaluation of the selectivity- and pair-aware components of the framework.

3.4. Ablation study of SPaR-GNN components

To assess the contribution of individual SPaR-GNN components, we performed a systematic ablation study with the scaffold-closure split using the same evaluation protocol and 100 random seeds. As summarized in Table 6, the full SPaR-GNN achieved the best mean performance across all seven pair-sensitive metrics, including a cliff RMSE of 2.083, a Spearman correlation of 0.646, and a direction accuracy of 0.742 and a strict-cliff RMSE of 2.813, a Spearman correlation of 0.592, and a direction accuracy of 0.880.

Among the evaluated components, direct selectivity supervision and pair-difference regression displayed the largest contributions. Removing direct selectivity supervision reduced cliff and strict-cliff Spearman

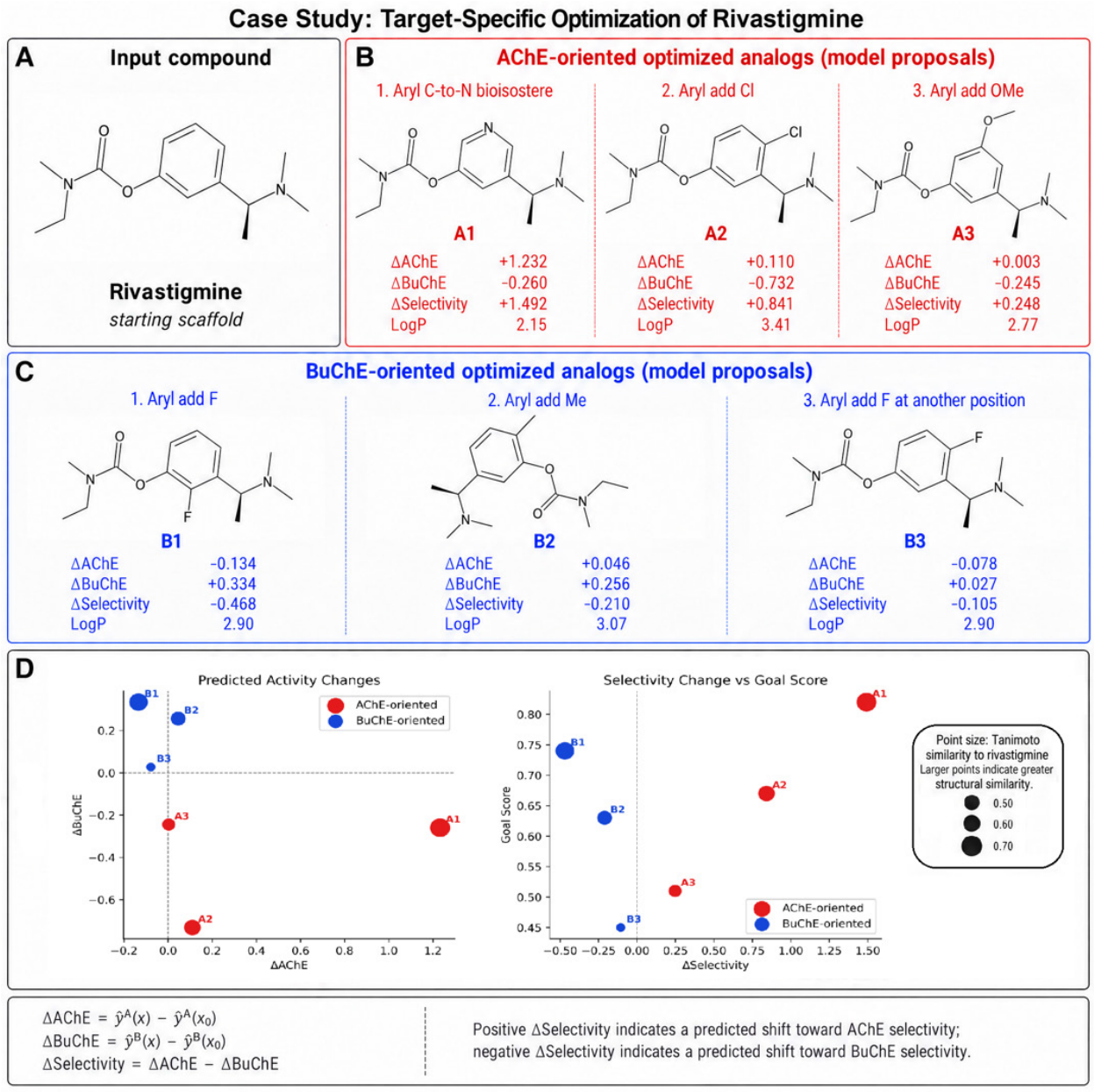


Figure 5. Optimization of compound rivastigmine. A, Rivastigmine was used as the input molecule for template-based local analog generation. **B,** Top-ranked candidates pursuant to the AChE-oriented optimization objective. **C,** Top-ranked candidates pursuant to the BuChE-oriented optimization objective. **D,** Distribution of prioritized candidates according to predicted changes in AChE and BuChE activity (left) and predicted selectivity change versus goal score (right). ΔAChE and ΔBuChE denote differences in predicted pIC_{50} relative to the input molecule, and ΔSelectivity = ΔAChE - ΔBuChE. Positive ΔSelectivity indicates a predicted shift toward AChE selectivity, whereas negative values indicate a predicted shift toward BuChE selectivity. Red and blue points represent candidates prioritized pursuant to the AChE- and BuChE-oriented objectives, respectively. Point size represents Tanimoto similarity to rivastigmine, with larger points indicating greater structural similarity.

correlations, as did removing pair-difference regression. Removal of the consistency, ranking, or cliff-weighting objectives also consistently degraded pair-sensitive performance, indicating complementary contributions from these components.

The full SPaR-GNN also substantially outperformed the independent single-task GINE, parameter-matched GINE, and Pairwise Morgan-XGB baselines, particularly in terms of rank consistency for the cliff and strict-cliff subsets. These results suggest that the performance gains are not attributable to model capacity or pairwise inputs alone but are associated with the combined contribution of the selectivity- and pair-aware objectives under scaffold-level generalization.

3.5. Goal-directed optimization of rivastigmine analogs

Rivastigmine was selected as a representative starting scaffold to demonstrate objective-dependent candidate generation and ranking under two predefined scoring configurations: AChE-oriented and BuChE-oriented objectives (Figure 5).

Under the AChE-oriented configuration, the highest-ranked analog was generated by an aryl C-to-N bioisosteric replacement and showed predicted changes of ΔAChE = +1.232, ΔBuChE = -0.260, and ΔSelectivity = +1.492. Aromatic chlorination and methoxy substitution also produced candidates with positive predicted selectivity shifts, although smaller than that of the top-ranked analog (Figure 5B).

Under the BuChE-oriented configuration, a different set of analogs was prioritized, including fluorine- and methyl-substituted candidates (Figure 5C). The highest-ranked candidate showed predicted changes of ΔBuChE = +0.334, ΔAChE = -0.134, and ΔSelectivity = -0.468. Other highly ranked candidates displayed similar model-predicted trends consistent with the specified objective.

The two scoring configurations produced clearly separated candidate rankings (Figure 5D). AChE-oriented candidates were characterized by positive predicted Δ AChE and selectivity shifts, whereas BuChE-oriented candidates tended toward positive predicted ΔBuChE and negative selectivity shifts, demonstrating objective-dependent prioritization from the same starting scaffold.

Top-ranked candidates retained moderate similarity to rivastigmine, with a Tanimoto similarity generally ranging from approximately 0.6 to 0.7, indicating local rather than extensive scaffold modification. This case study therefore serves as a demonstration of the workflow for objective-dependent candidate generation and prioritization. Because ranking was based on predictions from SPaR-GNN, these results do not constitute experimental evidence of improved activity, selectivity, pharmacokinetic properties, or therapeutic potential.

4. Discussion

Computational approaches for cholinesterase inhibitor discovery have conventionally focused on single-target activity, whereas AD lead optimization may require consideration of both AChE and BuChE activity and their relative selectivity. BuChE activity may become increasingly important during AD progression as AChE activity declines in later stages (*37*). We therefore developed SPaR-GNN to jointly model AChE activity, BuChE activity, and selectivity through multitask prediction combined with selectivity- and pair-aware supervision.

Results indicated that the main advantage of SPaR-GNN lies in selectivity modeling rather than isolated single-target prediction. Under scaffold-controlled evaluation, the fingerprint-based MLP remained competitive and outperformed the conventional GNN and GINE baselines on several single-molecule metrics, indicating that graph representation alone does not guarantee improved generalization to unseen chemotypes. SPaR-GNN likewise did not outperform conventional baselines on individual AChE or BuChE prediction but did achieve the best performance in terms of single-molecule selectivity. Together with the ablation results, this suggests that its advantage is more closely associated with the selectivity- and pair-aware learning framework than with the use of a graph architecture alone. This distinction is relevant to medicinal chemistry, where optimization often depends on shifting the relative activity between targets rather than maximizing either target independently.

The scaffold-closure split provided a stringent assessment of generalization by eliminating train-test overlap at the molecule, SMILES, scaffold, and pair-endpoint levels. This is particularly important for pair-sensitive evaluation because structural overlap can otherwise lead to optimistic estimates of performance (*10*). In this setting, SPaR-GNN displayed good pair-sensitive performance, substantiating its ability to prioritize selectivity changes among structurally related compounds beyond conventional pointwise regression.

The ablation study further corroborated the contribution of the combined framework. The full SPaR-GNN achieved the best mean performance across all evaluated pair-sensitive metrics, while removal of individual selectivity- or pair-aware objectives consistently reduced performance. Direct selectivity supervision and pair-difference regression produced particularly large effects on cliff and strict-cliff rank consistency, while consistency, ranking, and cliff-weighting objectives provided complementary contributions. The superiority of the full model over independent single-task GINE, parameter-matched GINE without pair-aware losses, and Pairwise Morgan-XGB further suggests that the observed gains arise from the combined selectivity- and pair-aware formulation rather than model capacity or pairwise input alone.

SPaR-GNN was further integrated into a goal-directed molecular optimization workflow combining template-based local editing, physicochemical filtering, goal-aware scoring, and candidate ranking. In a rivastigmine case study, AChE- and BuChE-oriented objectives produced distinct candidate rankings from the same starting scaffold while retaining relatively local structural modifications. Because candidate generation was restricted to a predefined template library and ranking relied on predictions from the same SPaR-GNN model, this analysis should be interpreted as a demonstration of the workflow for objective-dependent candidate generation and prioritization rather than evidence of improved pharmacological properties, experimentally validated structure-activity relationships, or superior compounds.

Nonetheless, there were several limitations. First, ChEMBL and BindingDB contain measurements obtained under heterogeneous experimental protocols and assay conditions (*17,18*). Although data standardization and quality control were implemented, residual variability may propagate into selectivity labels because selectivity is calculated based on both AChE and BuChE measurements. Experimentally harmonized dual-target datasets would therefore enhance future selectivity modeling. Second,

the proposed candidates are computational predictions and need to be validated experimentally. Third, the predefined medicinal chemistry template library improves interpretability and chemical plausibility but restricts accessible chemical space. Future work could expand this space using reaction-based generation, reinforcement learning, or large language model-guided molecular design (*38,39*).

Acknowledgements

The authors wish to thank the developers and maintainers of the ChEMBL and BindingDB databases for making their data publicly available.

Funding: This work was supported by a Grant-in-Aid from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (24K14216).

Conflict of Interest: The authors have no conflicts of interest to disclose.

References

1. Scheltens P, De Strooper B, Kivipelto M, Holstege H, Chételat G, Teunissen CE, Cummings J, van der Flier WM. Alzheimer's disease. Lancet. 2021; 397:1577-1590.
2. El Haj M, Kapogiannis D. Time distortions in Alzheimer's disease: A systematic review and theoretical integration. NPJ Aging Mech Dis. 2016; 2:16016.
3. DeTure MA, Dickson DW. The neuropathological diagnosis of Alzheimer's disease. Mol Neurodegener. 2019; 14:32.
4. Lei C, Zhang B, Yamaguchi J, Tamura R, Zhong W, Cao X, Liu Y, Seki M, Suzuki Y, Suzuki K, Tanida I, Uchiyama Y, Hisatsune T. Dual inhibition of IRAK1/TAK1 signaling in astrocytes reduces accelerated mortality in human APOE4 knock-in APPswe/PSEN1dE9/P301S-Tau triple transgenic mouse model. J Neuroinflammation. 2025; 22:237.
5. Melnikova I. Therapies for Alzheimer's disease. Nat Rev Drug Discov. 2007; 6:341-342.
6. Darvesh S. Butyrylcholinesterase as a diagnostic and therapeutic target for Alzheimer's disease. Curr Alzheimer Res. 2016; 13:1173-1177.
7. Gao H, Jiang Y, Zhan J, Sun Y. Pharmacophore-based drug design of AChE and BChE dual inhibitors as potential anti-Alzheimer's disease agents. Bioorg Chem. 2021; 114:105149.
8. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011; 162:1239-1249.
9. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018; 23:1241-1250.
10. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018; 9:513-530.
11. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019; 59:3370-3388.
12. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. Proceedings of the 34th International Conference on Machine Learning, PMLR. 2017; 70:1263-1272.
13. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, Li Z, Luo X, Chen K, Jiang H, Zheng M. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 2020; 63:8749-8760.
14. Hu X, Liu G, Zhao Y, Zhang H. Activity cliff-aware reinforcement learning for de novo drug design. J Cheminform. 2025; 17:54.
15. Liu S, Qu M, Zhang Z, Cai H, Tang J. Structured multi-task learning for molecular property prediction. Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, PMLR. 2022; 151:8906-8920.
16. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, Clevert DA, Hochreiter S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018; 9:5441-5451.
17. Mendez D, Gaulton A, Bento AP, *et al*. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 2019; 47:D930-D940.
18. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016; 44:D1045-D1053.
19. Landrum G. RDKit: Open-source cheminformatics; 2024. *https://www.rdkit.org/* (accessed June 1, 2026).
20. Lane RM, Potkin SG, Enz A. Targeting acetylcholinesterase and butyrylcholinesterase in dementia. Int J Neuropsychopharmacol. 2006; 9:101-124.
21. Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. J Med Chem. 2012; 55:2932-2942.
22. Han R, Yoon H, Kim G, Lee H, Lee Y. Revolutionizing medicinal chemistry: The application of artificial intelligence (AI) in early drug discovery. Pharmaceuticals (Basel). 2023; 16:1259.
23. Xu K, Hu W, Leskovec J, Jegelka S. How powerful are graph neural networks? arXiv preprint. 2018; https://doi.org/10.48550/arXiv.1810.00826.
24. Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, Metni H, van Hoesel C, Schopmans H, Sommer T, Friederich P. Graph neural networks for materials science and chemistry. Commun Mater. 2022; 3:93.
25. Mandi J, Bucarey V, Tchomba MMK, Guns T. Decision-focused learning: Through the lens of learning to rank. Proceedings of the 39th International Conference on Machine Learning, PMLR. 2022; 162:14935-14947.
26. Dablander M, Hanser T, Lambiotte R, Morris GM. Exploring QSAR models for activity-cliff prediction. J Cheminform. 2023; 15:47.
27. McNair D. Artificial intelligence and machine learning for lead-to-candidate decision-making and beyond. Annu Rev Pharmacol Toxicol. 2023; 63:77-97.
28. Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF. Generative models for molecular discovery: Recent advances and challenges. Wiley Interdiscip Rev Comput Mol Sci. 2022; 12:e1608.
29. Park J, Ahn J, Choi J, Kim J. Mol-AIR: Molecular reinforcement learning with adaptive intrinsic rewards for goal-directed molecular generation. J Chem Inf Model. 2025; 65:2283-2296.
30. Rogers D, Hahn M. Extended-connectivity fingerprints. J

Chem Inf Model. 2010; 50:742-754.
31. Gao W, Fu T, Sun J, Coley C. Sample efficiency matters: A benchmark for practical molecular optimization. Advances in Neural Information Processing Systems. 2022; 35:21342-21357.
32. Boström J, Brown DG, Young RJ, Keserü GM. Expanding the medicinal chemistry synthetic toolbox. Nat Rev Drug Discov. 2018; 17:709-727.
33. Roughley SD, Jordan AM. The medicinal chemist's toolbox: An analysis of reactions used in the pursuit of drug candidates. J Med Chem. 2011; 54:3451-3479.
34. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; p. 785-794.
35. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986; 323:533-536.
36. Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, Leskovec J. Strategies for pre-training graph neural networks. In: International Conference on Learning Representations. 2020; https://doi.org/10.48550/arXiv.1905.12265.
37. Greig NH, Lahiri DK, Sambamurti K. Butyrylcholinesterase: An important new target in Alzheimer's disease therapy. Int Psychogeriatr. 2002; 14 Suppl 1:77-91.
38. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. 2018; 555:604-610.
39. M Bran A, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024; 6:525-535.

Received August 19, 2026; Revised October 12, 2026; Accepted October 15, 2026.

[§]These authors contributed equally to this work.
**Address correspondence to:*
Kenji Karako, Department of Surgery, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan.
E-mail: tri.leafs@gmail.com

Released online in J-STAGE as advance publication October 20, 2026.