Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and Evaluation of Target-Specific Machine-Learning Scoring Functions for Monoamine Oxidase B

Sherif Adel Arafa Elsabbagh

sherif.elsabbagh90@gmail.com

Galala University https://orcid.org/0000-0002-4348-7723

Research Article

Keywords: Machine-learning scoring functions, structure-based virtual screening, MAOB, target-specific scoring functions, decoy bias, protein–ligand interaction fingerprints, benchmark design

Posted Date: September 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11070192/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Target-specific machine-learning scoring functions (MLSFs) have improved virtual screening for several targets, but their performance depends heavily on benchmark design and modeling choices. Here, we develop and evaluate target-specific MLSFs for human monoamine oxidase B (MAO-B), systematically comparing three modeling decisions: classification versus regression, structure-based (PLEC) versus combined structure- and ligand-based (PLEC + Morgan fingerprint) featurization, and untuned versus hyperparameter-tuned models. Generic scoring function performance was first established on the DUD-E benchmark, then re-evaluated on an independent test set composed exclusively of experimentally confirmed inactive compounds to control for property-matched decoy bias. On DUD-E, RF-Score-VS achieved the highest early enrichment among generic scoring functions (NEF1% = 0.789), but declined to NEF1% = 0.061 on the true-inactive test set — indistinguishable from the native Smina docking score and close to the random baseline (NEF1% = 0.024). Target-specific models trained on 2,328 actives and 121,764 DeepCoy-generated decoys consistently outperformed their classification counterparts: regression exceeded classification for all five algorithms tested. Combining Morgan fingerprints with PLEC features improved performance in five of ten baseline algorithm–task combinations. The best-performing model was tuned XGB regression with combined features (NEF1% = 0.245), exceeding the best generic scoring function on the same test set (CNN-Score, NEF1% = 0.082) by approximately threefold. Hyperparameter optimization did not reliably improve early-recognition performance, and a structural similarity analysis showed that the 25 test actives with above-threshold training-set similarity were retrieved at a substantially higher rate than the remaining 92 (28.0% vs. 5.4% for XGB; 28.0% vs. 2.2% for SVM), indicating that a meaningful share of early-recognition performance reflects structural analogy to known actives rather than generalization to novel chemical space. These findings argue for carefully designed benchmarks, explicit bias controls, and prospective validation in future target-specific MLSF studies.

1. Introduction

Structure-based virtual screening (SBVS) is widely used to identify active compounds from large chemical libraries when a three-dimensional model of the target is available (Li et al., 2020; Tran-Nguyen et al., 2023). In this workflow, docking generates candidate binding poses, and the success of screening depends heavily on the scoring function used to rank protein–ligand complexes (Ain et al., 2015; Wang et al., 2019). Classical scoring functions remain useful, but their simplified, often linear treatment of protein–ligand interactions limits their ability to capture complex binding patterns and has contributed to a plateau in virtual-screening accuracy (Ain et al., 2015; Li et al., 2020; Wójcikowski et al., 2017; Zhang et al., 2023).

The growing availability of protein structures, compound libraries, and integrated bioactivity data has created favorable conditions for machine-learning scoring functions (MLSFs) that can learn nonlinear relationships from large training sets (Ain et al., 2015; Ghislat et al., 2021; Qamar et al., 2022; Tran-Nguyen et al., 2023). Across retrospective and prospective studies, MLSFs have generally improved hit

rates, potency recovery, and virtual-screening performance relative to classical scoring approaches (Ain et al., 2015; Li et al., 2020; Tran-Nguyen et al., 2023). These gains have been especially notable for target-specific models, which are trained on data from a given protein or target family and often outperform generic scoring functions in practical screening settings (Qamar et al., 2022; Wang et al., 2019).

A robust target-specific MLSF therefore requires three elements: high-quality target-relevant data, informative representations of ligand and protein–ligand interactions, and an algorithm suited to the prediction task (Zhang et al., 2023). Recent pipelines have used descriptors and fingerprints such as MACCS, ECFP4/ECFP6, IFP, SIFP, and PLEC, together with algorithms including random forest, gradient boosting, deep neural networks, and graph convolutional networks, to improve target-specific screening across diverse systems (Junaid et al., 2025; Qamar et al., 2022; Wang et al., 2025; Zhang et al., 2023). Practical workflows for MLSF development now emphasize curated public activity data, target-specific train/test partitioning, and rigorous evaluation against existing baselines (Ballester, 2019; Tran-Nguyen et al., 2023).

At the same time, recent work shows that MLSF performance depends critically on how models are trained and evaluated. Generic MLSFs often generalize poorly to dissimilar targets (Shen et al., 2020). Even strong models lose performance under sequence- or pocket-based cross-validation, indicating that benchmark design can inflate apparent accuracy when analogue leakage is not controlled (Zhu et al., 2022). Likewise, on the relatively unbiased LIT-PCBA benchmark, target-specific MLSFs tended to outperform Glide SP but did not consistently exceed ligand-based 2D QSAR models, emphasizing that both accuracy and novelty should be assessed explicitly (Shen et al., 2021). These findings have motivated bias-aware practices such as similarity-constrained splitting, property-matched decoy generation, multi-pose augmentation, and consensus scoring (Ericksen et al., 2017; Junaid et al., 2025; McGibbon et al., 2022).

Recent target-specific MLSF studies have further shown that model performance depends strongly on modeling choices that are not yet fully understood. For PARP1, regression-based models trained on quantitative bioactivity data generally outperformed classification-based models trained on the same features, with the exception of random forest, for which the reverse was true (Caba et al., 2024). A similar advantage of regression over classification was independently observed for TRPM8, where switching from classification to regression increased early enrichment several-fold using otherwise identical features and training data (James & Ballester). The same study also demonstrated that training-set class imbalance must approximate the imbalance expected in a real screening library: augmenting an initially near-balanced training set with additional inactive compounds to match the test set's imbalance increased NEF1% several-fold, whereas ligand-only models lacking structural information did not benefit from the same adjustment.

Human monoamine oxidase B (MAO-B) is a therapeutically important mitochondrial flavin-dependent enzyme involved in the catabolism of dopamine and other biogenic amines (Cieślak et al., 2024; Murugan et al., 2020; Shrestha et al., 2026). Because MAO-B activity contributes to dopamine depletion,

oxidative stress, and neurotoxic metabolite formation, MAO-B inhibition remains an established strategy for managing Parkinson's disease and is also relevant to other neurodegenerative disorders. Although many MAO-B inhibitors with micromolar to nanomolar activity are already known, there is still a need for compounds with improved affinity, selectivity, and pharmacokinetic properties, particularly because MAO inhibitor design is complicated by isoform similarity and historical safety liabilities (Cieślak et al., 2024; Murugan et al., 2020; Shrestha et al., 2026).

For MAO-B, the case for a target-specific MLSF is particularly strong because standard force-field-based ranking methods have shown weak performance on this target. In a benchmark of 3,753 MAO-B-active compounds, AutoDock, AutoDock Vina, and MM-GBSA were unreliable or underperformed for inhibitor ranking, whereas machine-learning approaches, especially KNN, gave better ranking accuracy at lower computational cost (Murugan et al., 2020). Notably, this prior MAO-B work relied exclusively on ligand-based physicochemical descriptors, without docking or structural protein–ligand interaction information; a structure-based target-specific approach incorporating protein–ligand interaction fingerprints has not, to our knowledge, been applied to MAO-B.

Against this background, developing a human MAO-B-specific ML scoring function for virtual screening is well justified. MAO-B has clear therapeutic relevance, available bioactivity data, and documented limitations of classical scoring on this target (Murugan et al., 2020). At the same time, the broader MLSF literature supports the view that target-specific, carefully validated, and bias-aware models can improve screening efficiency and hit prioritization over generic docking scores alone. Accordingly, the present study develops and evaluates a target-specific machine-learning scoring function for human MAO-B, using protein–ligand interaction fingerprints (PLEC) extracted from docked poses. Because prior target-specific MLSF studies report inconsistent, target-dependent conclusions regarding classification versus regression modeling, the added value of combining structural and ligand-based feature representations, and the benefit of hyperparameter optimization, these three modeling choices are systematically compared for MAO-B, alongside benchmarking against established generic scoring functions.

2. Methods

2.1 Receptor preparation and molecular docking

The crystal structure of human monoamine oxidase B (PDB: 1S3B), including its bound FAD cofactor, was used as the docking receptor throughout this study. The structure was prepared using PyMOL (https://www.pymol.org/) and PDBFixer (Eastman et al., 2017): hydrogens were added using PDBFixer's standard protonation assignment at pH 7.4, the co-crystallized reference ligand was extracted separately from the protein coordinates, crystallographic water molecules were removed, missing residues were added using PDBFixer, and a single protein chain was retained.

All molecules were docked against a rigid receptor using Smina (Koes et al., 2013) with the default Smina scoring function, a search space centered on the co-crystallized reference ligand (autobox), a box

size of 30 × 30 × 30 Å, exhaustiveness = 1, a single output pose retained per molecule, and a fixed random seed (0) for reproducibility. These parameters were applied identically to every docking run performed in this study.

2.2 Virtual screening performance metrics

Virtual screening performance was quantified using the enrichment factor at the top 1% of the ranked library (EF1%), defined as the proportion of true active compounds recovered within the top 1% of a ranked molecule set, relative to the proportion expected under random ranking. EF1% values were additionally normalized (NEF1%) by the maximum EF1% attainable for each specific dataset, given its active:inactive ratio, enabling comparison across datasets of differing size and composition (Tran-Nguyen et al., 2023). NEF1% was selected as the primary performance metric, rather than ROC-AUC, BEDROC, or logAUC, for its direct focus on early recognition — the fraction of a screening library that is realistically prioritized for experimental testing — consistent with recent target-specific scoring function studies (Caba et al., 2024; James & Ballester). Precision–recall (PR) curves, computed across the full ranked molecule set, were used to visualize screening performance beyond the single top-1% cutoff.

2.3 Generic scoring function benchmarking

Before developing a target-specific scoring function, the performance of existing, off-the-shelf (generic) scoring functions was evaluated on MAO-B to establish a baseline for comparison, using the DUD-E benchmark (Mysinger et al., 2012) for this target (168 experimentally confirmed active compounds and 6,931 property-matched decoys; 7,099 molecules; active:inactive ratio ≈ 1:41.3), docked as described above.

Docked poses were rescored using four scoring functions: the native Smina docking score; CNN-Score (GNINA v1.3, run in score-only mode), an ensemble of convolutional neural networks trained on protein–ligand complex structures (McNutt et al., 2025); RF-Score-VS v2, a random forest model trained on the DUD-E collection across 102 targets, including MAO-B itself (DUD-E target code: aofb) (Wojcikowski et al., 2017); and an interaction-fingerprint-based similarity score (IFP) impletemeted in ODDT (Wojcikowski et al., 2015), computed as the Tanimoto similarity between each docked pose's protein–ligand interaction fingerprint and that of the co-crystallized reference ligand.

2.4 Bioactivity data curation

Bioactivity data for MAO-B were obtained from PubChem (UniProt P27338) and ChEMBL (CHEMBL2039, version 39, downloaded August 2026). Only half-maximal inhibitory concentration (IC50) measurements reported with a standard relation type of "=" (exact value) or ">" (censored) were retained, following Caba et al. (2024); measurements reported as Ki, Kd, or EC50, or with a "<" relation type, were excluded. Compound structures were standardized using RDKit 2026.03.6: salts were stripped, charges were neutralized, tautomers were canonicalized. Measurements were pooled across both data sources by standardized SMILES, yielding 7,798 compounds. For compounds with multiple pooled measurements, IC50 values were retained only if their standard deviation did not exceed 2 nM; 380 compounds were

excluded for exceeding this threshold. Compounds were classified as active if at least one exact measurement indicated an IC50 ≤ 1 μM, consistent with a threshold widely used in the field; as inactive if every available measurement supported inactivity (an exact measurement > 1 μM, or a censored measurement with a lower bound > 1 μM); and discarded otherwise (93 compounds, ambiguous or contradictory activity evidence). For regression modeling, active compounds were assigned a pIC50 calculated as −log10(mean IC50 in molar units) from their exact measurements. Following Caba et al. (2024), every inactive compound was assigned a fixed pIC50 of 2, with no claim made about the optimality of this choice. The final curated dataset comprised 2,454 active and 4,871 inactive compounds, each annotated with a binary activity classification and a pIC50 value, enabling classification and regression models to be trained and evaluated on an identical set of compounds throughout this study. By construction, all active compounds had a pIC50 ≥ 6 and all inactive compounds were assigned a fixed pIC50 of 2.

2.5 Decoy generation and test set construction

Property-matched decoys were generated for the 2,454 curated actives using DeepCoy (Imrie et al., 2021), at a ratio of 100 generated and 50 retained decoys per active. All molecules were docked as described in Section 2.1. Of 2,454 actives, 2,445 (99.6%) docked successfully and passed structural validation, and of 4,871 curated inactives, 4,828 (99.1%) docked successfully, and of 122,250 decoys, 122,114 (99.9%) docked successfully. Decoys corresponding to the nine excluded actives were removed from the decoy set (orphan filtering), yielding 121,764 decoys. Training and test sets were constructed as follows: property-matched decoys were confined entirely to the training set, and test-set inactives were drawn only from experimentally confirmed inactive compounds, isolating property-matched decoy bias. Active compounds were partitioned using Butina clustering (Butina, 1999) on Morgan fingerprints (radius 2, 2,048 bits; Tanimoto similarity threshold ≥ 0.70). Test-set actives were drawn exclusively from singleton clusters, preventing structurally similar actives from appearing in both sets; as a consequence, test-set actives are structurally distinct from training by construction, which may understate performance relative to a prospective screening library. Of 861 clusters identified, 360 were singletons; 117 singleton-cluster actives were selected at random to match the active:inactive ratio of the DUD-E benchmark (≈ 1:41.3), allowing comparison at the same early-recognition cutoff. The final training set comprised 2,328 actives and 121,764 decoys (124,092 molecules); the final test set comprised 117 actives and 4,828 inactives (4,945 molecules; active:inactive ratio 1:41.26).

2.6 Featurization

Protein–ligand extended connectivity (PLEC) interaction fingerprints were extracted from docked poses using ODDT, with fragment depths of 1 (ligand) and 5 (protein) and a fingerprint size of 4,092 bits. To evaluate the added value of ligand-based information, Morgan (extended-connectivity) fingerprints (radius 2, 2,048 bits), computed directly from each compound's structure using RDKit, were additionally generated and concatenated to the PLEC features, yielding a combined 6,140-feature representation (hereafter MF+PLEC).

2.7 Target Specific model development

Target-specific scoring functions were developed using five algorithms: random forest (RF), extreme gradient boosting (XGB), support vector machine (SVM), a single-hidden-layer artificial neural network (ANN), and a multi-layer deep neural network (DNN), implemented using scikit-learn (Pedregosa et al., 2011) and XGBoost (Chen & Guestrin, 2016). For each algorithm, both classification and regression models were trained, using the identical training and test sets and feature representations described above. Given the size of the training set (124,092 compounds), an exact support vector machine was found to be computationally intractable; SVM was instead implemented as a Nystroem kernel approximation (radial basis function kernel, 1,500 components) followed by stochastic gradient descent, using scikit-learn. Given the severe class imbalance of the training set (2,328 actives to 121,764 decoys, ≈ 1:52), class-weighted loss functions were used for classification models where supported (class_weight="balanced" for RF and SVM; scale_pos_weight, set to the training set's inactive:active ratio, for XGB). ANN and DNN models used early stopping on a held-out validation split. Full implementation details, including exact hyperparameter values, are available in the accompanying code repository. All final results are reported as the median across five repeated training runs with different random seeds.

2.8 Hyperparameter optimization

Hyperparameters were optimized using Optuna (Akiba et al., 2019), via ten trials of Bayesian search per algorithm and feature representation, each trial evaluated using five-fold cross-validation on the training set. Search spaces are listed in **Table S1**. For regression models, the cross-validation objective was the mean Spearman rank correlation between predicted and true pIC50 across folds; for classification models, the objective was the mean Average Precision across folds. Both objectives were chosen over alternatives such as raw error (MSE) or Matthews correlation coefficient (MCC) for their focus on ranking quality, consistent with this study's early-recognition-based evaluation metric (NEF1%). Following selection of the best-performing hyperparameter combination for each algorithm and feature representation, final models were trained and evaluated as described above.

2.9 Generic scoring function baseline on the true-inactive test set

To provide a baseline for comparison with target-specific models, the same three generic scoring functions evaluated in Section 2.3 [Smina, CNN-Score, RF-Score-VS] were re-evaluated by rescoring the docked poses of the final test set described in Section 2.5 (117 actives, 4,828 experimentally confirmed inactive compounds), rather than the DUD-E benchmark.

3. Results

3.1 Generic scoring function benchmarking

RF-Score-VS achieved the highest performance of the four scoring functions evaluated on the MAO-B/DUD-E benchmark (NEF1% = 0.789), recovering 56 of 168 true actives (33.3%) within the top 1% of the ranked library — four-fold higher than the second-best method, CNN-Score (NEF1% = 0.197). The native

Smina docking score performed considerably worse (NEF1% = 0.155), and the IFP score performed worst overall (NEF1% = 0.028). Precision–recall curves (Fig. 1) show that RF-Score-VS maintained precision near 1.0 out to a recall of approximately 0.2–0.3, whereas the other three scoring functions declined sharply at low recall.

Because MAO-B was among RF-Score-VS's own training targets, and because DUD-E decoys are property-matched rather than experimentally confirmed inactive compounds (Sieg et al., 2019), this result cannot be interpreted as a measure of generalization to a novel target or a realistic screening scenario. This is directly examined in the next section, where RF-Score-VS is re-evaluated against a test set composed exclusively of experimentally confirmed inactive compounds rather than property-matched decoys.

3.2 Generic scoring functions on the true-inactive test set

On the test set of 117 actives and 4,828 experimentally confirmed inactive compounds, all three generic scoring functions performed close to the random baseline (NEF1% = 0.024), substantially worse than on the DUD-E benchmark. RF-Score-VS declined from NEF1% = 0.789 on the DUD-E benchmark to NEF1% = 0.061 on this test set. CNN-Score achieved the highest performance among the three generic scoring functions (NEF1% = 0.082, approximately 3.4-fold above random), followed by Smina and RF-Score-VS (both NEF1% = 0.061, approximately 2.5-fold above random). These results indicate that generic scoring functions provide only modest early enrichment for MAO-B once evaluated against experimentally confirmed inactive compounds rather than property-matched decoys.

3.3 Regression versus classification, and the effect of hyperparameter optimization

Table S2 and Fig. 2 summarize the performance of all five algorithms, trained as both regression and classification models, using PLEC features, with baseline and hyperparameter-tuned configurations. Regression outperformed classification for all five algorithms under both baseline and tuned configurations. This difference was most pronounced for RF and DNN models, which produced NEF1% = 0.000 as classifiers at baseline (versus 0.102 and 0.143, respectively, as regressors). Hyperparameter optimization did not consistently improve performance. Of ten algorithm–task combinations, tuning improved NEF1% for four (XGB regression, ANN regression, RF classification, DNN classification), left four unchanged (RF regression, SVM regression, XGB classification, ANN classification), and reduced performance for two (DNN regression, SVM classification). This inconsistency mirrors prior reports that hyperparameter optimization guided by cross-validation does not reliably translate into improved early-recognition performance on an independent test set. Considering each algorithm–task combination at its best-performing configuration (baseline or tuned), six of ten exceeded the best generic scoring function (CNN-Score, NEF1% = 0.082): SVM and DNN regression at baseline (NEF1% = 0.143), XGB and ANN regression after tuning (NEF1% = 0.122), RF regression (NEF1% = 0.102 at both baseline and tuned), and SVM classification at baseline (NEF1% = 0.122).

3.4 Effect of combining structural and ligand-based features

Table S3 summarizes baseline and hyperparameter-tuned performance using the combined Morgan fingerprint and PLEC (MF + PLEC) representation. Figure 3 compares baseline performance directly between the two feature representations (PLEC alone, MF + PLEC) for each algorithm and modeling type. At baseline, combining Morgan fingerprints with PLEC features improved performance in five of ten algorithm–task combinations, left three unchanged, and reduced performance in two. The largest improvement was observed for XGB regression (0.061 → 0.143), followed by XGB classification (0.041 → 0.102), SVM regression (0.143 → 0.184), ANN regression (0.082 → 0.122), and DNN classification (0.000 → 0.020). Performance declined for ANN classification (0.041 → 0.020) and DNN regression (0.143 → 0.122), and remained unchanged for RF regression, RF classification, and SVM classification.

Hyperparameter tuning of the MF + PLEC representation improved performance in three combinations, left one unchanged, and reduced performance in six. In the tuned regression panel (Fig. 4, bottom right), XGB is clearly separated from all other curves, maintaining precision above 0.20 out to a recall of approximately 0.10 — the strongest sustained early-recognition behavior of any curve in the figure, consistent with its NEF1% of 0.245, a 1.7-fold increase over its own baseline (0.143). XGB regression (MF + PLEC, tuned) was the best-performing model in the entire study, exceeding the best generic scoring function (CNN-Score, NEF1% = 0.082) by 3.0-fold and the best PLEC-only result (SVM and DNN regression, baseline, NEF1% = 0.143) by 1.7-fold. The same tuning procedure produced the opposite outcome for SVM regression, the strongest MF + PLEC baseline model (NEF1% = 0.184): tuning reduced its performance to 0.082, visible in Fig. 4 as a marked decline between the baseline and tuned regression panels. This value coincides exactly with the best generic scoring function, and sits below every PLEC-only regression baseline, reinforcing the inconsistent effect of hyperparameter optimization observed with PLEC features alone (Section 3.3).

3.5 Structural similarity of retrieved hits

The true actives retrieved in the top 1% by the two best-performing models (XGB regression, MF+PLEC, tuned, NEF1% = 0.245; SVM regression, MF+PLEC, baseline, NEF1% = 0.184) were characterized by their maximum Tanimoto to any training-set molecule. XGB retrieved 12 true actives, 7 of which (58.3%) exceeded 0.70 similarity to a training-set molecule; SVM retrieved 9 true actives, 7 of which (77.8%) exceeded 0.70. In both cases, the nearest training-set neighbor was almost never an inactive compound (decoy) (0/12 for XGB; 1/9 for SVM), suggesting this similarity primarily reflects structural overlap with training actives rather than decoy bias. More broadly, 25 of 117 test-set actives (21.4%) had maximum similarity ≥ 0.70 to a training-set molecule, despite being drawn from singleton Butina clusters. These 25 structurally overlapping test actives were retrieved at a substantially higher rate than the remaining 92 (28.0% vs. 5.4% for XGB; 28.0% vs. 2.2% for SVM).

4. Discussion

The present results show that the apparent strength of generic scoring functions on MAO-B depends strongly on the evaluation set used. On the DUD-E benchmark, RF-Score-VS achieved the highest early enrichment among the generic methods examined, but this performance collapsed when the same method was re-evaluated on a test set composed exclusively of experimentally confirmed inactive compounds rather than property-matched decoys. This pattern is consistent with broader evidence that decoy-based benchmarks can inflate apparent screening performance and that bias control is essential when assessing machine-learning scoring functions for prospective use (McGibbon et al., 2022). The modest performance of Smina, CNN-Score, and RF-Score-VS on the true-inactive MAO-B test set illustrates the benefit of a target-specific approach here: whereas all three generic scoring functions performed only slightly above the random baseline on this harder test set, several MAO-B-specific models exceeded the best generic comparator. This agrees with the general view that target-specific scoring functions often outperform generic alternatives in practical virtual screening, particularly when target-relevant activity data are available (Tran-Nguyen et al., 2023). It is also especially plausible for MAO-B, where earlier work showed that conventional force-field-based ranking methods were unreliable or underperformed, while machine-learning approaches ranked MAO-B inhibitor complexes more effectively .

Regression was consistently more effective than classification for all five algorithms with PLEC features, and the contrast was strongest for RF and DNN, whose baseline classifiers failed to retrieve any active within the top 1% while their regression counterparts achieved measurable enrichment. This result aligns with recent target-specific MLSF studies showing that quantitative bioactivity labels can improve early recognition relative to binary active/inactive training (Caba et al., 2024; James & Ballester). A likely explanation is that regression preserves potency information within the active class, allowing the model to learn rank order among binders rather than only the decision boundary between actives and inactives. This difference was statistically significant (Welch's t-test, p = 0.0001; **Figure S1**)

Combining Morgan fingerprints with PLEC features produced gains that were positive overall but not universal. Adding ligand-based fingerprints improved five of ten baseline algorithm–task combinations, with the largest gains observed for XGB regression and SVM regression, while two combinations declined and three were unchanged (not statistically significant when pooled across all combinations, p = 0.695; **Figure S1**. This mixed pattern is consistent with the broader literature: richer or complementary descriptors can improve virtual screening by capturing information that coarse interaction summaries miss, but the benefit depends on how well the representation matches the learning algorithm and the target (Junaid et al., 2025; Tahir Ul Qamar et al., 2022). The best-performing model in the present work, tuned XGB regression with MF+PLEC features, likely benefited from this complementary encoding of ligand chemotype and docked interaction pattern. Hyperparameter optimization did not produce uniform gains. We found improvement in only four of ten PLEC-based comparisons and in only three of ten MF+PLEC comparisons, with several models unchanged and others worsened after tuning (not statistically significant when pooled, p = 0.464; **Figure S1**). That result cautions against assuming that

better cross-validation objectives necessarily translate into better early recognition on an external test set, especially under class imbalance and chemically constrained train-test.

The structural similarity analysis indicates that a meaningful share of early-recognition performance reflects structural analogy to training data rather than generalization to novel chemical space. Both best-performing models retrieved the 25 test actives with above-threshold training-set similarity at substantially higher rates than the remaining 92. Two contributing factors should be distinguished. First, test actives were drawn from singleton Butina clusters as a deliberate strategy to reduce training–test structural overlap, but Butina clustering is greedy and order-dependent, so singleton membership does not guarantee pairwise dissimilarity to the full training set. Second, even in the absence of this effect, a model trained on known MAO-B actives would be expected to prioritize chemotypes resembling those actives. This does not undermine the finding that the MAO-B-specific models outperform the generic scoring functions on the same test set, but it does temper the interpretation of model performance as evidence of generalization to structurally novel chemistry.

Several limitations should temper interpretation of our findings. First, all compounds were docked to a single rigid MAO-B structure with low exhaustiveness and a single retained pose, so errors in pose generation or receptor conformation directly constrain the quality of the PLEC representation used for learning. Second, regression assigned every inactive compound a fixed pIC50 of 2, which enabled identical compound sets to be used for classification and regression but remains a modeling convention rather than a validated biochemical truth. Third, the test-set actives were drawn from singleton clusters, which was intended to strengthen structural independence from training data but, as shown in Section 3.5, did not fully achieve this. Fourth, class imbalance was corrected for RF, SVM, and XGB models but not for ANN and DNN, this may partly explain why the neural models underperformed the tree-based and kernel-based approaches. Fifth, final results are reported as the median of five training–test runs, a lower number of repeats than in some closely related studies. Finally, this is a retrospective study; no prospective validation was performed.

5. Conclusion

This study developed and evaluated target-specific machine-learning scoring functions for human MAO-B, systematically comparing three modeling choices: classification versus regression, PLEC versus combined Morgan+PLEC features, and untuned versus hyperparameter-tuned models. Three results stand out.First, the apparent strength of generic scoring functions on MAO-B depends critically on the evaluation benchmark. RF-Score-VS, the strongest performer on the DUD-E benchmark (NEF1% = 0.789), declined to NEF1% = 0.061 when re-evaluated against experimentally confirmed inactive compounds — indistinguishable from the native Smina docking score. This illustrates the limited realism of decoy-based benchmarks and the value of true-inactive test sets for evaluating scoring functions intended for prospective use. Second, regression consistently outperformed classification for all five algorithms tested, and combining structural (PLEC) with ligand-based (Morgan) features improved performance for most algorithm–task combinations. The best-performing model overall was tuned XGB regression with

MF+PLEC features (NEF1% = 0.245), exceeding the best generic scoring function on the same test set (CNN-Score, NEF1% = 0.082) by approximately threefold. Third, hyperparameter optimization did not reliably improve early-recognition performance, and a structural similarity analysis showed that a substantial share of early-recognition gains reflected structural analogy to known actives rather than generalization to novel chemical space. These findings argue for carefully designed benchmarks, explicit bias controls, and prospective validation in future target-specific scoring function studies for MAO-B and related targets.

Declarations

Data availability

All bioactivity data used in this study were obtained from publicly available databases: PubChem (UniProt P27338) and ChEMBL (CHEMBL2039, version 39, downloaded August 2026). The DUD-E MAO-B benchmark is publicly available from the DUD-E database. All code used during this strudy, curated datasets, and all intermediate files generated during this study are available at [https://github.com/sherifaelsabbagh/MAOB-MLSF].

Competing interests

The author declares no competing interests.

References

1. Ain Q, Aleksandrova A, Roessler F, Ballester P (2015) Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. Wiley Interdiscip Rev Comput Mol Sci 5:405–424. https://doi.org/10.1002/wcms.1225
2. Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) *Optuna: A Next-generation Hyperparameter Optimization Framework* Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA. https://doi.org/10.1145/3292500.3330701
3. Ballester P (2019) Selecting machine-learning scoring functions for structure-based virtual screening. Drug discovery today Technol 32–33:81–87. https://doi.org/10.1016/j.ddtec.2020.09.001
4. Butina D (1999) J Chem Inf Comput Sci 39(4):747–750. https://doi.org/10.1021/ci9803381. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets
5. Caba K, Tran-Nguyen VK, Rahman T, Ballester PJ (2024) Comprehensive machine learning boosts structure-based virtual screening for PARP1 inhibitors. J Cheminform 16(1):40. https://doi.org/10.1186/s13321-024-00832-1

6. Chen T, Guestrin C (2016) *XGBoost: A Scalable Tree Boosting System* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785
7. Cieślak M, Danel T, Krzysztyńska-Kuleta O, Kalinowska-Tłuścik J (2024) Machine learning accelerates pharmacophore-based virtual screening of MAO inhibitors. *Scientific Reports*, *14*. https://doi.org/10.1038/s41598-024-58122-7
8. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang LP, Simmonett AC, Harrigan MP, Stern CD, Wiewiora RP, Brooks BR, Pande VS (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol 13(7):e1005659. https://doi.org/10.1371/journal.pcbi.1005659
9. Ericksen S, Wu H, Zhang H, Michael L, Newton M, Hoffmann F, Wildman S (2017) Machine Learning Consensus Scoring Improves Performance Across Targets in Structure-Based Virtual Screening. J Chem Inf Model 57:1579–1590. https://doi.org/10.1021/acs.jcim.7b00153
10. Ghislat G, Rahman T, Ballester P (2021) Recent progress on the prospective application of machine learning to structure-based virtual screening. Curr Opin Chem Biol 65:28–34. https://doi.org/10.1016/j.cbpa.2021.04.009
11. Imrie F, Bradley AR, Deane CM (2021) Generating property-matched decoy molecules using deep learning. Bioinformatics 37(15):2134–2141. https://doi.org/10.1093/bioinformatics/btab080
12. James N, Ballester PJ TRPM8 Virtual Screening: The Essential Role of Target-Specific, Inactive-Enriched Machine-Learning Scoring Functions. ChemRxiv, *2026*(0609). https://doi.org/10.26434/chemrxiv.15004552/v1
13. Junaid M, Zeeshan M, Li W (2025) AttentionScore: A Target-Specific, Bias-Aware Scoring Function for Structure-Based Virtual Screening: A Case Study on METTL3. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.5c02142
14. Koes DR, Baumgartner MP, Camacho CJ (2013) Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J Chem Inf Model 53(8):1893–1904. https://doi.org/10.1021/ci300604z
15. Li H, Sze K-H, Lu G, Ballester P (2020) Machine-learning scoring functions for structure-based virtual screening. Wiley Interdisciplinary Reviews: Comput Mol Sci 11. https://doi.org/10.1002/wcms.1478
16. McGibbon M, Money-Kyrle S, Blay V, Houston D (2022) SCORCH: Improving structure-based virtual screening with machine learning classifiers, data augmentation, and uncertainty estimation. J Adv Res. https://doi.org/10.1016/j.jare.2022.07.001
17. McNutt AT, Li Y, Meli R, Aggarwal R, Koes DR (2025) GNINA 1.3: the next increment in molecular docking with deep learning. J Cheminform 17(1):28. https://doi.org/10.1186/s13321-025-00973-x
18. Murugan N, Muvva C, Jeyarajpandian C, Jeyakanthan J, Subramanian V (2020) Performance of Force-Field- and Machine Learning-Based Scoring Functions in Ranking MAO-B Protein–Inhibitor Complexes in Relevance to Developing Parkinson's Therapeutics. Int J Mol Sci 21. https://doi.org/10.3390/ijms21207648

19. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 55(14):6582–6594. https://doi.org/10.1021/jm300687e
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12(85):2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html
21. Qamar MTU, Zhu X-T, Chen L, Alhussain L, Alshiekheid M, Theyab A, Algahtani M (2022) Target-Specific Machine Learning Scoring Function Improved Structure-Based Virtual Screening Performance for SARS-CoV-2 Drugs Development. *International Journal of Molecular Sciences*, *23*. https://doi.org/10.3390/ijms231911003
22. Shen C, Hu Y, Wang Z, Zhang X, Pang J, Wang G, Zhong H, Xu L, Cao D, Hou T (2020) Beware of the generic machine learning-based scoring functions in structure-based virtual screening. Brief Bioinform. https://doi.org/10.1093/bib/bbaa070
23. Shen C, Weng G, Zhang X, Leung E, Yao X, Pang J, Chai X, Li D, Wang E, Cao D, Hou T (2021) Accuracy or novelty: what can we gain from target-specific machine-learning-based scoring functions in virtual screening? Brief Bioinform. https://doi.org/10.1093/bib/bbaa410
24. Shrestha R, Parajuli N, Chhetri SP, Tamang A, Poudel M, Shrestha MCS, Shrestha A, Bharati T, Maharjan S, Marasini B, B., Subin A (2026) Machine Learning Assisted Exploration of Natural Monoamine Oxidase-B Inhibitors Through Molecular Docking, Structural Dynamics and Pharmacokinetic Evaluation. *Advanced Theory and Simulations*, *9*. https://doi.org/10.1002/adts.202502222
25. Sieg J, Flachsenberg F, Rarey M (2019) In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. J Chem Inf Model 59(3):947–961. https://doi.org/10.1021/acs.jcim.8b00712
26. Tahir Ul Qamar M, Zhu XT, Chen LL, Alhussain L, Alshiekheid MA, Theyab A, Algahtani M (2022) Target-Specific Machine Learning Scoring Function Improved Structure-Based Virtual Screening Performance for SARS-CoV-2 Drugs Development. Int J Mol Sci 23(19). https://doi.org/10.3390/ijms231911003
27. Tran-Nguyen V, Junaid M, Simeon S, Ballester P (2023) A practical guide to machine-learning scoring for structure-based virtual screening. Nat Protoc 18:3460–3511. https://doi.org/10.1038/s41596-023-00885-w
28. Wang B, Junaid M, Li W (2025) Graph convolutional neural networks improved target-specific scoring functions for cGAS and kRAS in virtual screening. Comput Struct Biotechnol J 27:2176–2185. https://doi.org/10.1016/j.csbj.2025.05.023
29. Wang D, Cui C, Ding X, Xiong Z, Zheng M, Luo X, Jiang H, Chen K (2019) Improving the Virtual Screening Ability of Target-Specific Scoring Functions Using Deep Learning Methods. Front Pharmacol 10. https://doi.org/10.3389/fphar.2019.00924

30. Wójcikowski M, Ballester P, Siedlecki P (2017) Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, *7*. https://doi.org/10.1038/srep46710
31. Wojcikowski M, Ballester PJ, Siedlecki P (2017) Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep 7:46710. https://doi.org/10.1038/srep46710
32. Wojcikowski M, Zielenkiewicz P, Siedlecki P (2015) Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. J Cheminform 7:26. https://doi.org/10.1186/s13321-015-0078-2
33. Zhang X, Shen C, Jiang D, Zhang J, Ye Q, Xu L, Hou T, Pan P, Kang Y (2023) TB-IECS: an accurate machine learning-based scoring function for virtual screening. J Cheminform 15. https://doi.org/10.1186/s13321-023-00731-x
34. Zhu H, Yang J, Huang N (2022) Assessment of the Generalization Abilities of Machine-Learning Scoring Functions for Structure-Based Virtual Screening. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.2c01149

Figures

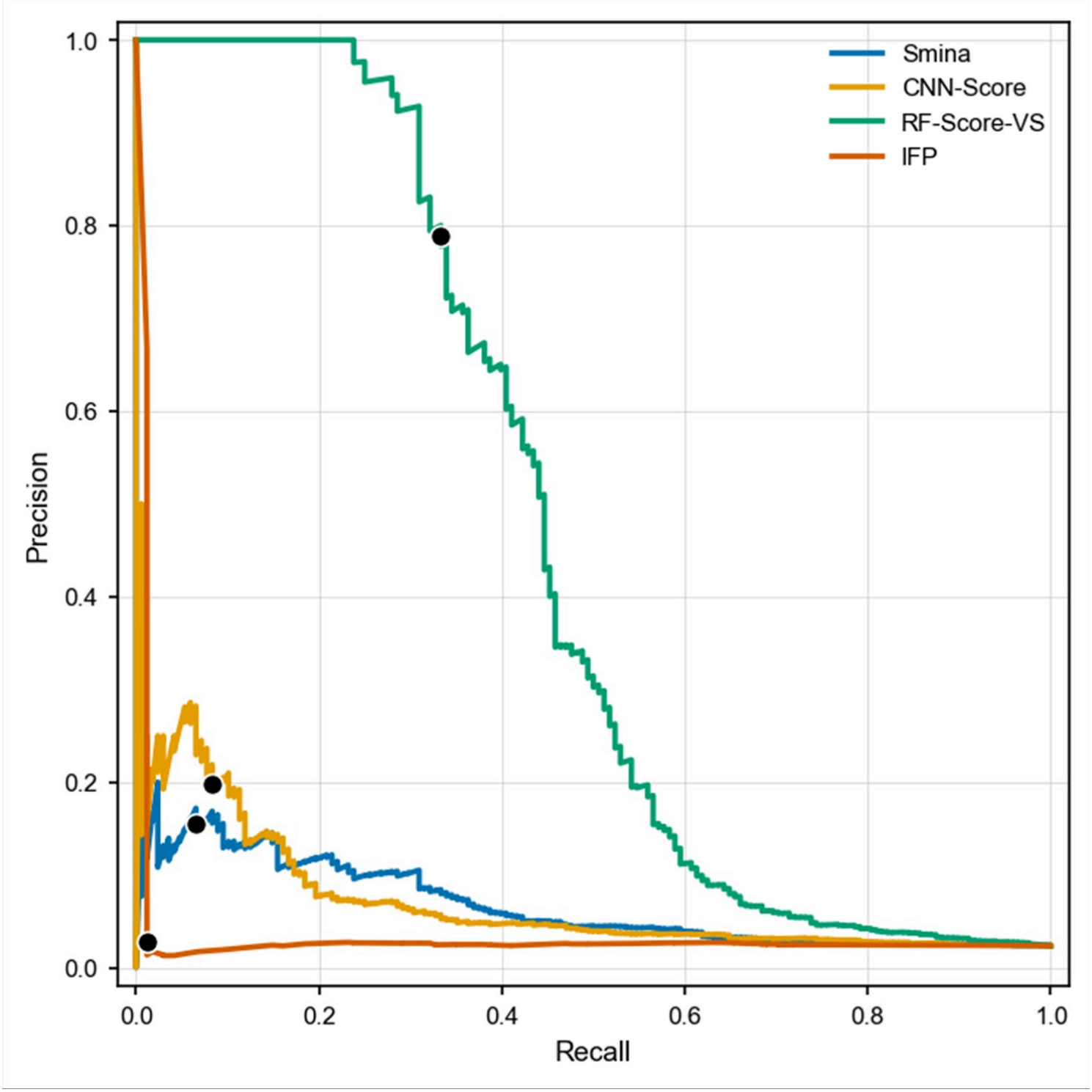


Figure 1

PR curves for four generic scoring functions rescoring docked poses of the MAO-B/DUD-E benchmark. The black markers represent the PR values corresponding to the top 1%-ranked molecules

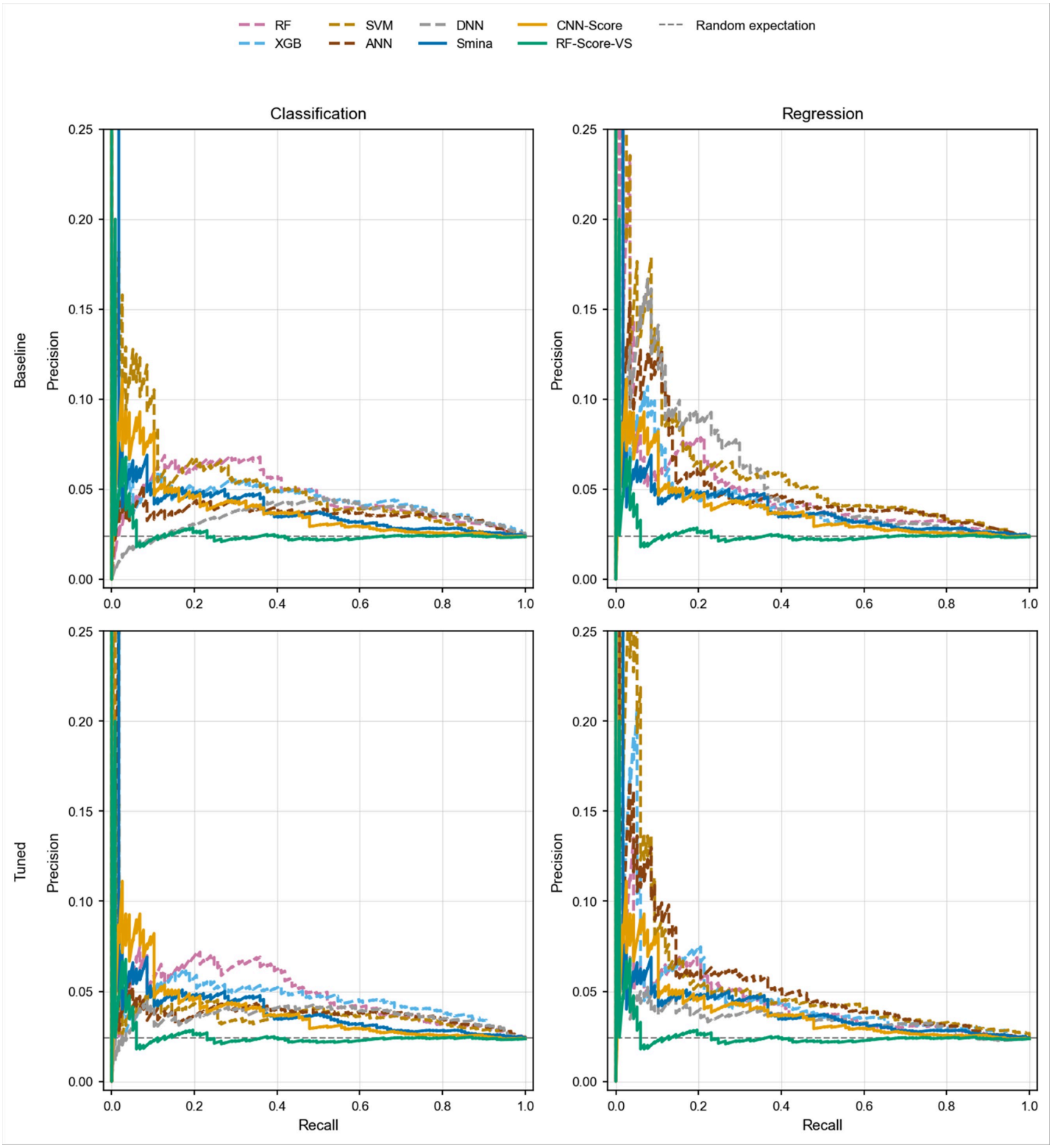


Figure 2

PRcurves for target-specific and generic scoring functions on the MAO-B test set, using PLEC features. Panels show baseline and hyperparameter-tuned configurations for classification (left) and regression (right) models. The y-axis is limited to 0–0.25 for visibility; the random baseline (NEF1% = 0.024) is shown as a horizontal dashed grey line. Each curve corresponds to the run giving the median NEF1% across five repeats.

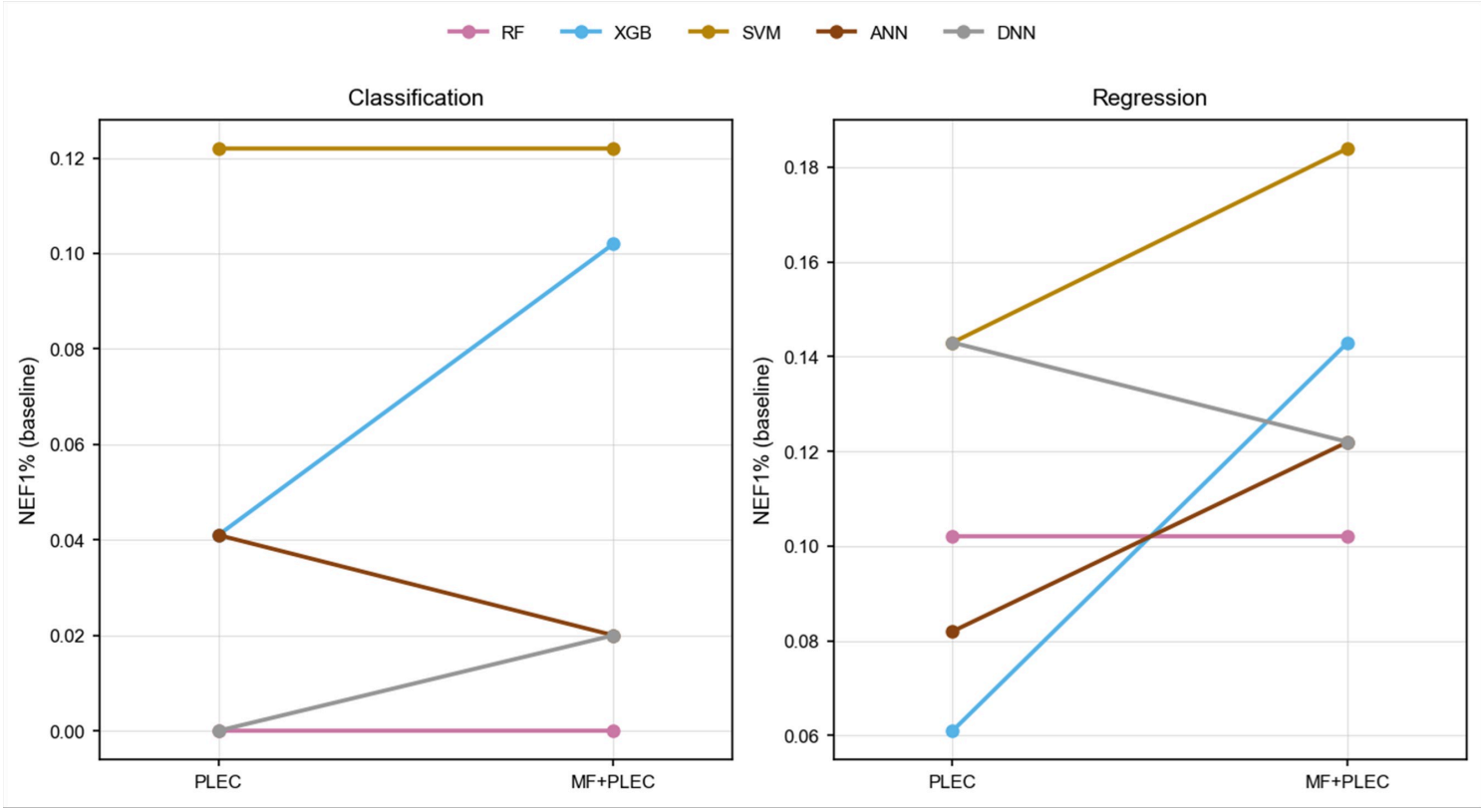


Figure 3

Baseline NEF1% for each of the five target-specific algorithms using PLEC features alone versus MF+PLEC representation. Panels show classification (left) and regression (right) models. Each line connects one algorithm's PLEC and MF+PLEC baseline values.

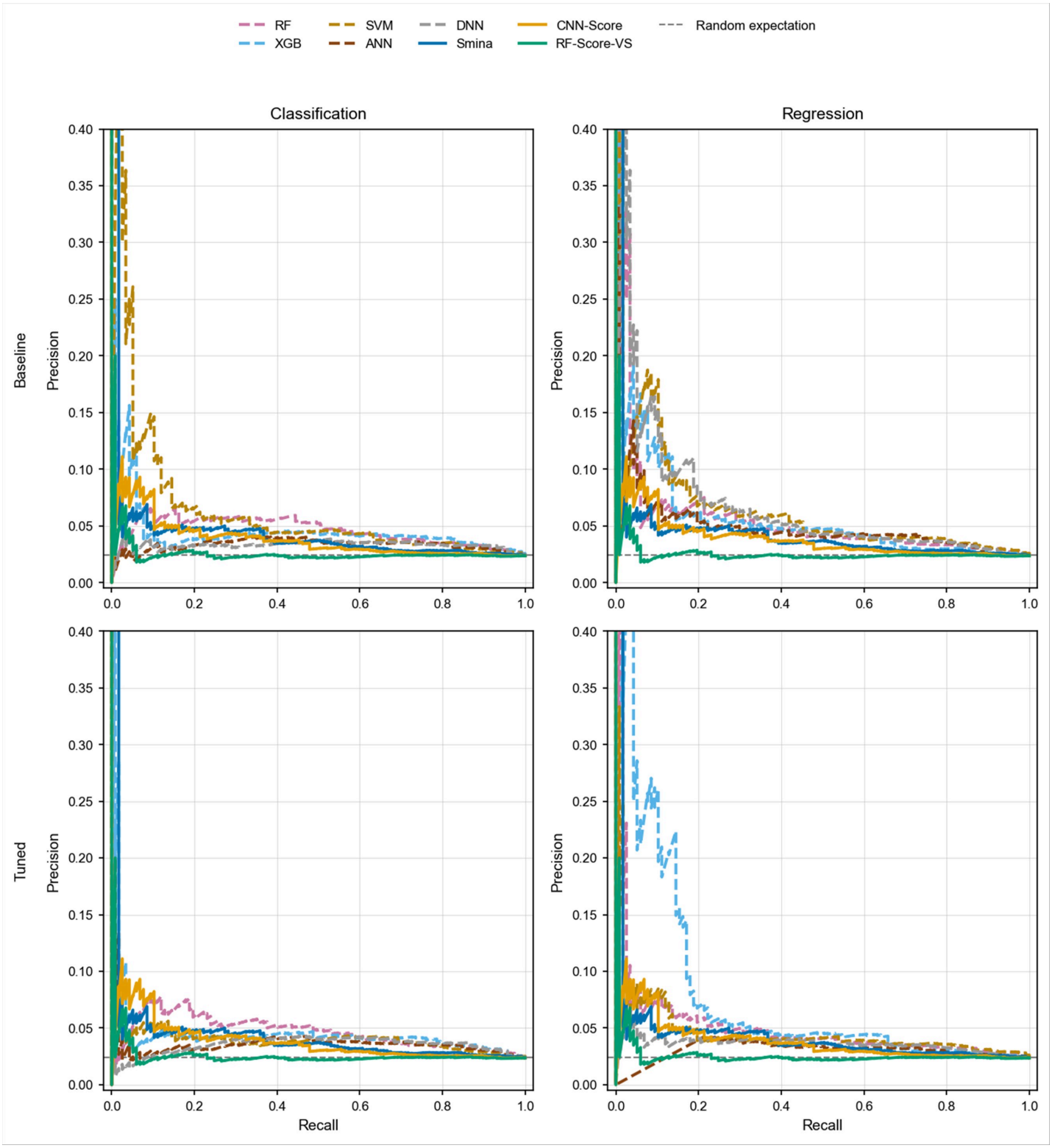


Figure 4

PR curves for target-specific MAO-B scoring functions using combined Morgan fingerprint and PLEC (MF+PLEC) features, and generic scoring functions on the true-inactive test set. Panels show baseline and hyperparameter-tuned configurations for classification (left) and regression (right) models. The random baseline (NEF1% = 0.024) is shown as a horizontal dashed grey line. Each curve corresponds to the run giving the median NEF1% across five repeats.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableFigures.docx