pharmaceutics

MDPI

Article

Multi-Task Learning of Blood–Brain Barrier Permeability and hERG Cardiotoxicity: A Random- and Scaffold-Split Evaluation on Experimental Data

Surya Veerendra Prabhakar Vattikuti [1,*], Saurabh Tiwari [2,*] and Priya Sharma [3]

[1] School of Mechanical Engineering, College of Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea
[2] School of Materials Science and Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea
[3] School of Pharmacy and Emerging Sciences, Baddi University of Emerging Sciences & Technology, Solan 173205, India; bhpriya02@gmail.com
* Correspondence: drprabu@ynu.ac.kr (S.V.P.V.); saurabht@yu.ac.kr (S.T.)

Abstract

Background/Objectives: Blood–brain barrier (BBB) permeability and cardiac hERG liability share physicochemical determinants, motivating the hypothesis that multi-task learning could improve BBB prediction. **Methods:** We tested this hypothesis using B3DB logBB regression ($n = 1058$) and a curated hERG dataset ($n = 8960$). A shared-encoder multi-task neural network was compared with a matched single-task network and Bayesian-optimized gradient boosting under ten random splits, 30 Murcko scaffold splits, and held-out classification sets. **Results:** Multi-task learning did not improve random-split R^2 (0.485 ± 0.016 vs. 0.482 ± 0.015). Under scaffold splitting, its mean R^2 was higher (-0.054 vs. -0.817) only because its failures were less severe (the same number of partitions had negative R^2); medians were nearly identical (0.394 vs. 0.398). Neither loss weighting, network tuning, nor removal of 39.6% of test compounds sharing a parent structure with hERG training data changed this. Gradient boosting was more robust (scaffold $R^2 = 0.388 \pm 0.188$; better in 28–29 of 30 partitions), although not significantly so under a dependence-corrected test. **Conclusions:** Network failures occurred only when a scaffold group containing 18 atypical compounds (e.g., xenon) entered the test set; after removing 20 such compounds, multi-task R^2 (0.409) matched gradient boosting (0.420). On 3031 held-out parent structures, AUCs were similar (0.764–0.772).

Keywords: blood–brain barrier; hERG cardiotoxicity; multi-task learning; neural network robustness; applicability domain; SHAP; Mordred descriptors

Check for updates

Academic Editor: Kyriakos Kachrimanis

Received: 8 September 2026
Revised: 23 October 2026
Accepted: 5 November 2026
Published: 9 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Blood–brain barrier (BBB) permeability remains a central determinant of the success of central nervous system (CNS) drug candidates. The broader importance of understanding central nervous system pharmacology is further highlighted by cerebrospinal fluid multi-omics studies identifying associations between lipid metabolism and functional outcomes following intracerebral hemorrhage [1]. Machine-learning approaches using curated two-dimensional (2D) molecular descriptors have demonstrated useful predictive performance for experimental logBB data [2–4]. However, evaluation under Murcko scaffold splitting, which more closely approximates prospective scaffold-hopping applications than random splitting, has revealed substantially lower and more variable performance [3,4]. These

observations emphasize the importance of assessing not only predictive accuracy but also robustness to structural distribution shifts.

A long-standing question in computational absorption, distribution, metabolism, excretion, and toxicity (ADMET) modelling is whether the prediction of one endpoint can be improved through multi-task learning by jointly training a shared representation on a mechanistically related endpoint for which more data are available [5]. BBB permeability and inhibition of the cardiac human Ether-à-go-go-Related Gene (hERG) potassium channel are both influenced by lipophilicity, basicity, and molecular polarity [6,7]. Moreover, hERG inhibition datasets assembled from ChEMBL and related sources are substantially larger than the experimentally curated B3DB logBB dataset, making hERG an attractive auxiliary task under the multi-task learning rationale that a larger, related dataset may regularize and inform a smaller target task [5]. Cardiovascular pharmacology also involves diverse mechanisms beyond ion-channel inhibition. Experimental studies have investigated the protective effects of calcitonin gene-related peptide on cardiac fibroblast senescence through Klotho expression [8] and the inhibition of TGF-β1-induced cardiac fibroblast activation by xanthohumol through PTEN/Akt/mTOR signaling [9]. These findings provide broader context for cardiovascular drug development, although neither mechanism directly establishes hERG inhibition liability.

Current BBB-permeability modelling has moved from physicochemical rules towards machine-learning models trained on curated benchmarks such as B3DB, which consolidated roughly 7800 classified and 1050 logBB-annotated compounds [10] and towards graph neural networks that learn molecular representations end-to-end [11,12]. In parallel, benchmarks on tabular data indicate that tree ensembles remain competitive with and frequently outperform neural networks at small to moderate sample sizes [13], which is relevant for logBB datasets of approximately 1000 compounds. Multi-task learning has been reported to help ADME-Tox prediction when large datasets are available [14], but its benefit is endpoint-dependent [6]. Whether a larger safety endpoint, such as hERG, can improve a small logBB model and how such a comparison should be evaluated when the two datasets share compounds have not been examined systematically.

This study directly tests whether auxiliary hERG supervision improves BBB permeability prediction. hERG was chosen because it is the largest openly available safety endpoint that shares physicochemical drivers with brain exposure (lipophilicity and basicity [5,6]); we also measured the relatedness of the two endpoints directly on the compounds present in both datasets. More directly coupled endpoints (P-glycoprotein efflux, plasma protein binding) were not tested. Relative to our earlier descriptor-based studies [2–4], which addressed single-endpoint logBB modelling, the present work differs in four respects: (i) it asks whether a second endpoint can improve the BBB model through a shared representation; (ii) it quantifies task-relatedness and the structural overlap between the auxiliary training data and the BBB test compounds, including per-partition overlap under scaffold splitting; (iii) it tests the robustness of the null result to loss weighting, neural-network tuning, and overlap removal; and (iv) it dissects the failure mode of neural networks under scaffold splitting, with and without compositionally atypical compounds. A Bayesian-optimized gradient-boosting model was included as a descriptor-based baseline, and SHAP analysis was retained only as a brief consistency check of that baseline (Section 3.7). This study is relevant to pharmaceutical development because early triage of CNS candidates for brain exposure and cardiac liability depends on such models, and their failure modes determine where they can be trusted.

2. Materials and Methods

2.1. Blood–Brain Barrier Dataset

The BBB regression dataset was obtained from the B3DB experimental database [10] and comprised 1058 compounds with logBB values ranging from −2.69 to +1.70 (mean = −0.078, SD = 0.752). The 41 Mordred two-dimensional descriptors defined previously [2] were taken from the published supplementary descriptor matrix [2], which is provided imputed and contains no missing values. The previously established train/validation/test partition (n = 740/106/212, stratified by logBB quintile) was retained for random-split analyses [2].

2.2. hERG Cardiotoxicity Dataset

hERG inhibition data were obtained from the publicly archived dataset accompanying Siramshetty et al. [15], comprising 8993 compounds (training set, n = 8154; independent validation set, n = 839) with binary hERG activity labels (10 µM potency threshold) integrated from three sources: NCATS Tox21 thallium-flux assay (n = 4508), ChEMBL25 bioactivity records (n = 3589), and their overlap (n = 57). After deduplication by canonical SMILES and removal of one compound that failed RDKit (RDKit, 2026.08.12) valence sanitization (a silicon-containing structure with non-standard valence), the working dataset comprised 8960 compounds (training, n = 8153; validation, n = 807), with a class balance of 6753 inactive/2207 active (24.6% active) in the full set. The same 41 descriptors were computed for the hERG SMILES using Mordred v1.2.0 [16] and RDKit; missing descriptor values (281/8960 compounds, predominantly Kier indices) were imputed with hERG training-set medians.

2.3. Model Architectures and Training

A shared dense encoder (41 → 256 → 128 → 64 units, batch normalization, ReLU activation, dropout = 0.3 at each layer) fed two task-specific heads: a BBB regression head (64 → 32 → 1, mean-squared-error loss) and an hERG classification head (64 → 32 → 1, binary cross-entropy loss with logits). The multi-task network was trained by alternating mini-batches from the BBB training set (batch size 32) and the hERG training set (batch size 128) within each epoch. In the primary analyses the combined loss was the unweighted sum of the two task losses, chosen as a neutral default; because the mean-squared error of logBB and the binary cross-entropy of the hERG label differ in scale, a 1:1 weighting does not guarantee equal influence on training, and a loss-weight sensitivity analysis (hERG weight 0.1, 0.5, and 2.0 with BBB weight 1) is therefore reported in Section 3.4. A single-task network with identical encoder capacity, trained only on the BBB regression objective, served as the matched baseline. Both networks used the Adam optimizer (learning rate 1×10^{-3}, weight decay 1×10^{-5}), early stopping on BBB validation loss (patience = 30 epochs, maximum 300 epochs), and RobustScaler feature standardization (median/IQR-based; BBB features scaled with a scaler fit on the BBB training partition only, hERG features with a scaler fit on the hERG training set). RobustScaler was chosen a priori as an outlier-resistant alternative to Z-score standardization; later, we show that it did not prevent the instability that was observed.

Gradient-boosting hyperparameters (scikit-learn GradientBoostingRegressor [17]) were optimized by Bayesian search (scikit-optimize v0.10.2, Gaussian-process surrogate, 50 evaluations, 10 random initial points, random seed 42). The search covered four hyperparameters: number of estimators (50–500), learning rate (0.01–0.3, log-uniform prior), maximum tree depth (2–8), and subsample fraction (0.5–1.0), using 5-fold cross-validated mean squared error on the random-split training partition (n = 740) as the objective. The optimal configuration (249 estimators, learning rate 0.0268, maximum depth 8, subsample

fraction 0.535; 5-fold CV $R^2 = 0.555 \pm 0.088$) was held fixed across the ten random-split seeds, thirty scaffold-split partitions, and all external sets. A simpler, non-optimized configuration (300 estimators, learning rate 0.05, maximum depth 5, subsample 0.8) was also evaluated for comparison. The comparison is therefore between a Bayesian-optimized gradient-boosting model and neural networks with a fixed, previously chosen configuration. To test whether this difference in tuning effort explains the gap, a limited random search over 20 neural-network configurations (including the default) was run with varying hidden-layer sizes ((256, 128, 64), (128, 64), (512, 256, 128), (128, 64, 32), (256, 128)), dropout (0.1–0.4), learning rate (3×10^{-4}–10^{-2}), weight decay (10^{-6}–10^{-3}), and BBB batch size (16, 32, 64), using 3-fold cross-validated MSE on the random-split training partition; the selected configuration was applied to both single- and multi-task networks.

2.4. Data Partitioning and Structural Overlap

Random-split evaluation reused the fixed, previously published B3DB training/validation/test partition (Section 2.1) for the BBB task and the native training/validation partition of the hERG dataset (Section 2.2) for the hERG task. For the neural networks, random weight initialization and mini-batch ordering varied across ten seeds (42, 1–9); for gradient boosting, the same ten seeds controlled the stochastic row subsampling used at each boosting iteration (subsample fraction < 1), which is the source of the seed-to-seed variability of gradient boosting despite its otherwise deterministic tree construction. Scaffold-split evaluation used Bemis–Murcko scaffolds (RDKit) [18], with scaffold groups (rather than individual compounds) randomly assigned to train (70%), validate (10%), and test (20%) subsets independently for each of 30 seeds (0–29), guaranteeing zero scaffold overlap between training and test in every partition (verified programmatically for each split). Thirty partitions, generated with different random seeds, were used to sample the variability of scaffold assignment. Because all partitions were resampling of the same 1058 compounds, they were not independent observations; therefore, comparisons across partitions were interpreted through effect sizes and a dependence-corrected test rather than through *p*-values alone. The random-split analysis retained ten seeds because it showed low variance. The hERG training/validation partition was held fixed across scaffold-split runs because hERG served only as an auxiliary task.

Because the hERG dataset was assembled independently of B3DB, compound overlap between the two datasets was determined in two ways: by exact canonical-SMILES match and by parent structure (largest fragment retained, charges neutralized, stereochemistry ignored; first block of the InChIKey). By canonical SMILES, 279 of the 1058 B3DB compounds (26.4%) also appeared in the hERG dataset, and 63 of the 212 random-split test compounds (29.7%) were present in the hERG training set that trains the shared encoder. By parent structure, the corresponding numbers were 397 (37.5%) and 84 of 212 (39.6%). Under scaffold splitting, the number of test compounds ($n = 213$ per partition) sharing a parent structure with the hERG training set ranged from 63 to 104 (29.6–48.8%; mean 39.1%) across the 30 partitions (by exact SMILES: 40–84; 18.8–39.4%; mean 28.8%). Therefore, the multi-task network had indirect exposure, via the auxiliary hERG task, to structural information about a substantial fraction of the compounds nominally held out for BBB evaluation. This does not constitute label leakage (hERG activity and logBB are independently measured quantities), but it means that the scaffold-split test sets are not independent of the auxiliary task. Because overlap could favor beneficial transfer, it cannot weaken a null result; its influence was nevertheless tested directly by retraining the multi-task network with every hERG training compound present in the corresponding BBB test fold removed (Sections 2.8 and 3.4).

2.5. SHAP Interpretability Analysis

To assess whether the Bayesian-optimized gradient-boosting model remained mechanistically consistent with established BBB permeability determinants, SHAP (SHapley Additive exPlanations) values were computed using TreeExplainer from the SHAP package (v0.54.1) [19]. TreeExplainer was applied to the seed-42 random-split model and 212 held-out test compounds. Global feature importance was quantified as the mean absolute SHAP value across the test set. As an independent model-agnostic cross-check, permutation importance (30 repeats, scikit-learn v1.9.1) was computed for the same model and test set.

2.6. Held-Out Classification Evaluation

Two held-out sets from B3DB were used. They came from the same resource as the training data and were therefore described as held-out rather than fully independent. Parent structures were defined as in Section 2.4. (i) The original external classification set ($n = 175$; 171 BBB^+, 4 BBB^-) [2] provides categorical labels without logBB values or a stated cut-off; 22 of its compounds share a parent structure with the regression set and were removed for the main analysis ($n = 153$; 151 BBB^+, 2 BBB^-); the analysis of all 175 compounds is given in the Supplementary Materials. (ii) An expanded set comprised the B3DB classification compounds whose parent structure was absent from the 1058-compound regression set, with each parent structure counted once ($n = 3031$; 1651 BBB^+, and 1380 BBB^-); the unfiltered classification set contained 6771 non-regression records (by canonical SMILES), many of which were stereoisomers, salts, or duplicates of one another. Labels derive from B3DB: a logBB threshold of −1 is listed for a subset of records, and the remainder derive from categorical literature assignments, so label definitions are not fully homogeneous. Models trained on the random-split training partition were applied without retraining, and predictions were averaged over the ten seeds (42, 1–9). Continuous logBB predictions served as ranking score for AUC-ROC, reported with 95% confidence intervals from 2000 bootstrap resamples of compounds and paired bootstrap differences between models. For threshold-dependent metrics, we report confusion counts and balanced accuracy at logBB = −1 (the B3DB convention for logBB-derived labels, corresponding to a brain-to-blood ratio of roughly 10%) and at logBB = 0 for continuity with the earlier analysis; neither threshold is guaranteed to correspond to the labelling criterion of every record. Because the multi-task network was trained with hERG data, results are also reported separately for compounds whose parent structure is ($n = 1047$) or is not ($n = 1984$) present in the hERG training set. The minimum heavy-atom count of the 175-compound set is 7, so it contains none of the compositionally atypical compounds of Section 3.5; under the same structural rule, 3 of the 3031 compounds of the expanded set (0.1%) qualify as atypical.

2.7. Statistical Analysis

Performance is reported as mean ± standard deviation across seeds or partitions. Because partitions are resampled from the same data, they are not independent, and the results are described by the mean paired difference with a bootstrap 95% confidence interval (10,000 resamples of the paired differences), the standardized effect size (Cohen's dz = mean difference/SD of differences), and the number of partitions in which one model was better. Paired two-sided *t*-tests and Wilcoxon signed-rank tests (non-parametric, appropriate for outlier-prone scaffold-split R^2 distributions) are reported for reference. As the *t*-test assumes independent observations, we also applied the corrected resampled *t*-test of Nadeau and Bengio [20], in which the variance of the paired differences is multiplied by (1/n + n_test/n_train) with n_test/n_train = 213/740. This correction was derived for random resampling and is used here as a conservative heuristic for scaffold-group partitions;

therefore, the conclusions emphasize the magnitude and consistency of the differences. Significance was set at $\alpha = 0.05$.

2.8. Sensitivity Analyses

Four additional sensitivity analyses were conducted using the same computational pipeline as the main results, each with ten random seeds and 30 scaffold partitions: (i) task-relatedness, as the correlation between logBB and the hERG label on compounds present in both datasets; (ii) overlap removal, retraining the multi-task network after removing all compounds present in the BBB test fold of each split from hERG training (by parent structure; exact-SMILES removal is also evaluated); (iii) loss-weight sensitivity (hERG weight 0.1, 0.5, 2.0); and (iv) cleaned-dataset retraining of all models after removing the 20 compositionally atypical compounds (fewer than four heavy atoms or no carbon atom). The exclusion rule is structural, not performance-based; it removes compounds that lie outside the drug-like organic chemical space for which a descriptor-based BBB model is intended (noble gases, lithium, small gases and solvents, SF_6). We state openly that the rule was formulated after the failure mode was first observed (Section 3.5), but it was applied uniformly to all 1058 compounds without reference to any model error. The analyses were performed using Python 3.13, scikit-learn 1.9.1, PyTorch 2.14.1, SHAP 0.54.1, RDKit, and Mordred 1.2.0.

3. Results

3.1. Random-Split Performance

Under the original random data partition, the Bayesian-optimized gradient-boosting model achieved test $R^2 = 0.581 \pm 0.014$ (RMSE = 0.488 ± 0.008, MAE = 0.333 ± 0.005 log units; 5-fold CV $R^2 = 0.555 \pm 0.088$ on the training partition), approaching the previously reported optimized result ($R^2 = 0.6043$ [2]). The non-optimized configuration yielded the same random-split performance (0.582 ± 0.011). The single-task neural network achieved $R^2 = 0.482 \pm 0.015$, and the multi-task network $R^2 = 0.485 \pm 0.016$ (Tables 1 and S1, Figure 1A). The single seed-42 run used for the SHAP and predicted-versus-actual analyses achieved R2 = 0.594 (Figure 2, Table S1). Gradient boosting was better than both networks in all ten seeds (mean difference +0.098, 95% CI +0.085 to +0.113, dz = 4.1 versus the single-task network; +0.096, 95% CI +0.082 to +0.109, dz = 4.1 versus the multi-task network; corrected $p = 1.1 \times 10^{-4}$ and 1.0×10^{-4}—Tables 2 and S2). The difference between the multi- and single-task networks was negligible (mean difference +0.002, 95% CI −0.009 to +0.013; dz = 0.13; paired *t*-test $p = 0.69$, Wilcoxon $p = 0.77$; multi-task better in 6 of 10 seeds). The hERG head of the multi-task network achieved AUROC = 0.828 ± 0.010 and AUPRC = 0.237 ± 0.022 on the held-out hERG validation set ($n = 807$; Figure 3), in line with published hERG models trained on comparable data (balanced accuracy 0.75–0.80 reported for the original consensus model on this dataset [13]), confirming that the shared encoder learned a useful hERG-relevant representation even though this did not transfer to improved BBB prediction.

Table 1. Summary performance across data-split regimes (mean ± SD).

Split	Model	BBB Test R^2	RMSE (log Units)	MAE (log Units)
Random (n = 740/106/212; 10 seeds)	Gradient boosting (optimized)	0.581 ± 0.014	0.488 ± 0.008	0.333 ± 0.005
	Single-task NN	0.482 ± 0.015	0.542 ± 0.008	0.410 ± 0.009
	Multi-task NN (BBB + hERG)	0.485 ± 0.016	0.541 ± 0.008	0.409 ± 0.008
Murcko scaffold (30 partitions; 70/10/20)	Gradient boosting (optimized)	0.388 ± 0.188	0.547 ± 0.049	0.404 ± 0.036
	Single-task NN	−0.817 ± 2.412	0.802 ± 0.414	0.480 ± 0.084
	Multi-task NN (BBB + hERG)	−0.054 ± 0.787	0.678 ± 0.180	0.457 ± 0.048

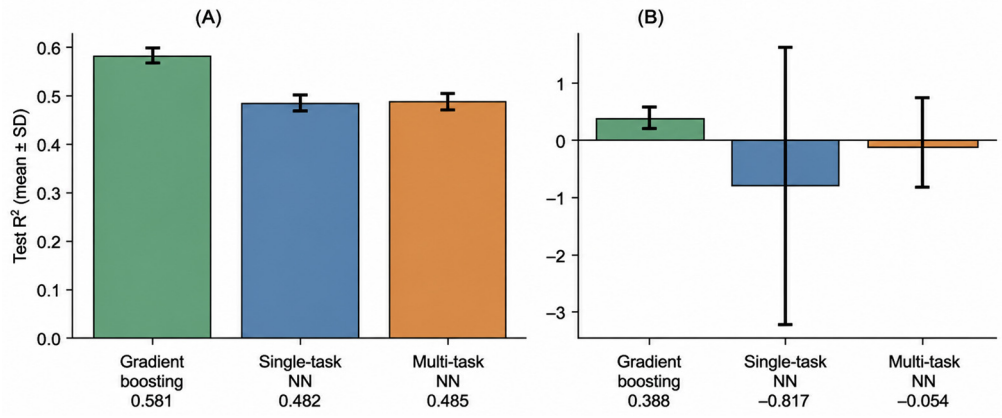

Figure 1. Test R^2 (mean ± SD) for Bayesian-optimized gradient boosting, single-task neural network, and multi-task (BBB + hERG) neural networks. (**A**) Random split (n = 1058; 10 seeds). (**B**) Murcko scaffold split (30 partitions).

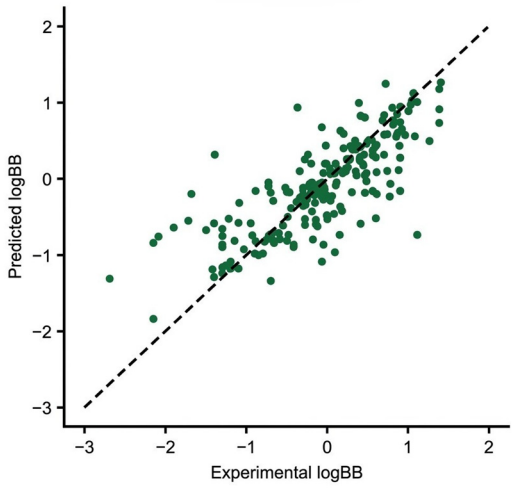

Figure 2. Predicted versus actual logBB, Bayesian-optimized gradient boosting, random-split held-out test set (seed 42, n = 212, R^2 = 0.594).

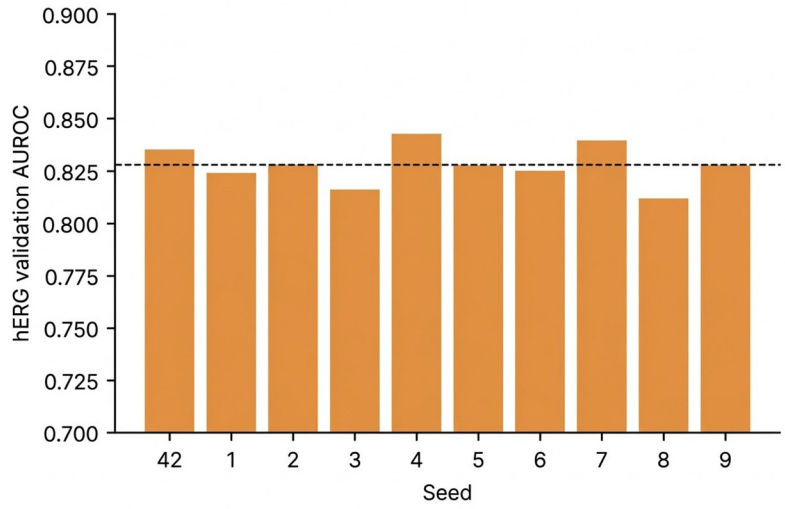

Figure 3. hERG classification AUROC of the multi-task network's hERG head across ten seeds and held-out hERG validation set (n = 807).

Table 2. Paired differences in test R^2 (mean difference with bootstrap 95% CI, standardized effect size dz, number of seeds/partitions in which the first model was better, and *p*-values from paired *t*-test, Wilcoxon test, and Nadeau–Bengio corrected test).

Split	Comparison	Mean Δ [95% CI]	dz	Better in	*p* (t)	*p* (Wilcoxon)	*p* (Corrected)
Random (10)	GB − single-task	+0.098 [0.085, 0.113]	4.07	10/10	4.2×10^{-7}	0.0020	1.1×10^{-4}
Random (10)	GB − multi-task	+0.096 [0.082, 0.109]	4.08	10/10	4.1×10^{-7}	0.0020	1.0×10^{-4}
Random (10)	Multi − single	+0.002 [−0.009, 0.013]	0.13	6/10	0.69	0.77	0.84
Scaffold (30)	GB − single-task	+1.205 [0.481, 2.080]	0.53	29/30	0.0074	1.3×10^{-8}	0.36
Scaffold (30)	GB − multi-task	+0.442 [0.239, 0.679]	0.70	28/30	0.0006	2.6×10^{-8}	0.23
Scaffold (30)	Multi − single	+0.763 [0.152, 1.530]	0.39	16/30	0.043	0.24	0.50

3.2. Scaffold-Split Performance and Model Robustness

Under Murcko scaffold-split evaluation across 30 partitions, the Bayesian-optimized gradient-boosting configuration achieved $R^2 = 0.388 \pm 0.188$ (median 0.482; range −0.02 to 0.60), with one of 30 partitions (3.3%) yielding a negative R^2. This was lower and more variable than the simpler, non-optimized configuration (300 estimators, learning rate 0.05, maximum depth 5, subsample 0.8), which achieved $R^2 = 0.440 \pm 0.120$ (median 0.464) with no negative partitions (difference −0.052, 95% CI −0.100 to −0.011; paired *t*-test $p = 0.037$, Wilcoxon $p = 0.34$, corrected $p = 0.49$). For context, a scaffold-split R^2 of 0.418 ± 0.085 was reported for LightGBM on B3DB [3]. The single-task network produced a negative test R^2 in 10 of 30 partitions (33.3%; range −8.94 to +0.54; median 0.398) and the multi-task network in 10 of 30 partitions (range −2.09 to +0.52; median 0.394), yielding mean R^2 of -0.817 ± 2.412 and -0.054 ± 0.787, respectively (Table 1, Figures 1B and 4). Gradient boosting achieved a higher R^2 than the single-task network in 29 of 30 partitions (mean difference +1.205, 95% CI +0.481 to +2.080; dz = 0.53) and then the multi-task network in 28 of 30 partitions (+0.442, 95% CI +0.239 to +0.679; dz = 0.70) (Table 2). The Wilcoxon test supported these differences ($p = 1.3 \times 10^{-8}$ and 2.6×10^{-8}), and the paired *t*-test was nominally significant ($p = 0.0074$ and 0.0006), but neither accounted for the dependence between partitions; under the Nadeau–Bengio corrected test, the differences were not significant ($p = 0.36$ and 0.23). Therefore, we describe the advantage of gradient boosting under scaffold splitting as consistent and of moderate to large size, concentrated in the lower tail of the neural-network distributions (medians 0.482 versus 0.398 and 0.394), rather than as statistically established.

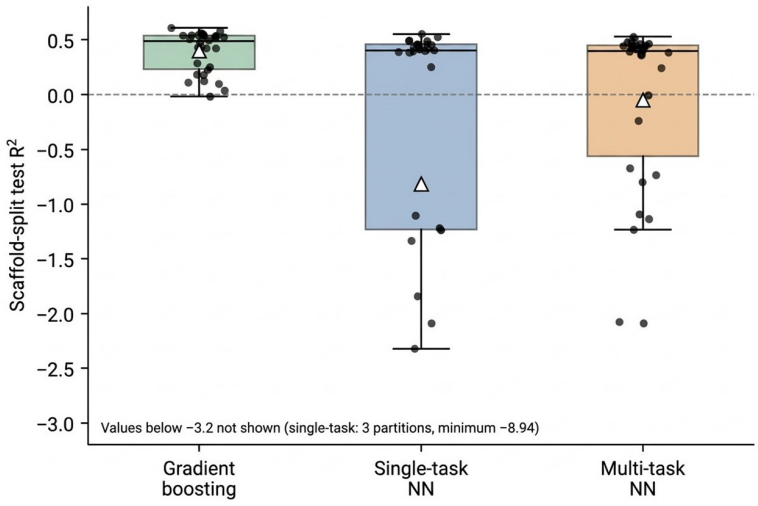

Figure 4. Test R^2 distributions for Bayesian-optimized gradient boosting, single-task NN, and multi-task NN across 30 Murcko scaffold partitions (box = interquartile range, line = median, triangle = mean, and points = individual partitions).

The multi-task network had a higher mean scaffold-split R^2 than the single-task network (mean difference +0.763, 95% CI +0.152 to +1.530; paired *t*-test $p = 0.043$), but the difference was carried by less severe failures of the multi-task network (negative R^2 in 10 of 30 partitions for both networks, but minimum −2.09 versus −8.94), not by better typical performance: the medians are nearly identical (0.394 versus 0.398), the multi-task network was better in only 16 of 30 partitions, and the difference was supported neither by the Wilcoxon test ($p = 0.24$) nor by the corrected test ($p = 0.50$). This stabilizing effect is investigated in Sections 3.4 and 3.5.

3.3. Task Relatedness and Structural Overlap

Among the 279 compounds present in both datasets by exact SMILES (55 hERG-active, 224 inactive), logBB correlated positively with the hERG label (Spearman $\rho = 0.32$, $p = 7 \times 10^{-8}$; point-biserial r = 0.32); logBB discriminated hERG-active from inactive compounds with AUC = 0.73, and mean logBB was 0.42 for hERG-active and −0.15 for inactive compounds (Table 3). Matching by parent structure gave 397 shared compounds (83 hERG-active) and the same relatedness ($\rho = 0.32$, $p = 3 \times 10^{-11}$; AUC = 0.73). Lipophilicity and basicity were associated with both endpoints (SLogP: $\rho = 0.45$ with hERG activity and 0.32 with logBB. nBase: 0.53 and 0.26), whereas TopoPSA, the dominant BBB descriptor, was unrelated to hERG activity in this subset ($\rho = -0.08$, $p = 0.18$) and only weakly related in the full hERG set ($\rho = -0.13$). The two endpoints are therefore moderately related through lipophilicity and basicity but not through polarity. This subset consists of compounds measured for both endpoints and is not representative of either dataset. Structural overlap with the hERG training set was substantial in every split (Section 2.4): 39.6% of the random-split test compounds and 29.6–48.8% (mean 39.1%) of the scaffold-partition test compounds, so scaffold-split test sets were not independent of the auxiliary task (Table S3).

Table 3. Relatedness of logBB and hERG activity of the 279 compounds present in both datasets (exact SMILES match; 55 hERG-active).

Quantity	Value
Spearman ρ (logBB, hERG label)	0.316 ($p = 7.0 \times 10^{-8}$)
AUC of logBB for hERG-active vs. inactive	0.729
Mean logBB, hERG-active/inactive	0.422/−0.151
ρ with hERG label/with logBB:SLogP	0.454/0.323
ρ with hERG label/with logBB:nBase	0.531/0.263
ρ with hERG label/with logBB:TopoPSA	−0.081/−0.562
ρ with hERG label/with logBB:MW	0.401/−0.113
Parent-structure match: shared compounds/ρ/AUC	397/0.325/0.731

3.4. Robustness of the Multi-Task Conclusion

Removing from hERG training all compounds sharing a parent structure with the BBB test fold (84 compounds on the random split; 63–104 per scaffold partition) did not change the results (Table 4). The random-split R^2 was 0.490 ± 0.019 with overlap removed, 0.485 ± 0.016 with overlap retained, and 0.482 ± 0.015 for the single-task network; removing the overlap changed R^2 by +0.005 (95% CI −0.006 to +0.017), and the removal-versus-single-task difference was +0.008 (95% CI −0.009 to +0.024). Under scaffold splitting, the median R^2 was 0.399, 0.394, and 0.398, respectively; the mean was −0.129 ± 1.062 with overlap removed (difference to the multi-task network with overlap: −0.076, 95% CI −0.286 to +0.119; paired *t*-test $p = 0.48$) (Table S4). Removing exact-SMILES duplicates only gave the same picture (Table S5). Therefore, overlap does not explain the null result. Varying the hERG loss weight (0.1, 0.5, 2.0) gave random-split R^2 of 0.479 ± 0.010, 0.490 ± 0.012,

and 0.483 ± 0.019 versus 0.482 for the single-task network (all differences within ±0.008; all 95% CIs include zero) and scaffold-split medians of 0.385, 0.403, and 0.401 versus 0.398 (Tbales 4 and S6). The neural-network random search reduced the cross-validated MSE from 0.276 (default) to 0.269 (hidden layers 512-256-128, dropout 0.4, learning rate 3×10^{-4}, weight decay 10^{-5}, batch size 32) (Table S7). With this configuration, random-split R^2 was 0.491 ± 0.012 (single-task) and 0.497 ± 0.021 (multi-task; difference +0.007, 95% CI −0.009 to +0.023), against 0.581 ± 0.014 for gradient boosting (differences +0.090, 95% CI +0.076 to +0.103, and +0.083, 95% CI +0.067 to +0.098). Under scaffold splitting, the tuned networks reached −0.377 ± 1.650 and −0.012 ± 0.817 (medians 0.383 and 0.399) versus 0.388 ± 0.188 (median 0.482) for gradient boosting, which was better in 27 of 30 partitions in both comparisons (corrected $p = 0.39$ and 0.31). Thus, within this limited tuning budget, tuning narrowed neither the random-split gap nor the scaffold-split median gap materially, and no loss weight, tuning setting, or overlap removal revealed a multi-task benefit.

Table 4. Sensitivity of multi-task conclusion (mean ± SD; median in brackets for scaffold; negative-R^2 partitions in parentheses).

Model/Condition	Random Split R^2 (10 Seeds)	Scaffold R^2 (30 Partitions)
Single-task NN (default)	0.482 ± 0.015	−0.817 ± 2.412 [0.398] (10 neg.)
Multi-task NN, hERG weight 1.0 (default)	0.485 ± 0.016	−0.054 ± 0.787 [0.394] (10 neg.)
Multi-task, hERG weight 0.1	0.479 ± 0.010	−0.053 ± 0.954 [0.385] (8 neg.)
Multi-task, hERG weight 0.5	0.490 ± 0.012	−0.153 ± 1.003 [0.403] (9 neg.)
Multi-task, hERG weight 2.0	0.483 ± 0.019	−0.093 ± 0.849 [0.401] (10 neg.)
Multi-task, overlap removed (parent structure)	0.490 ± 0.019	−0.129 ± 1.062 [0.399] (9 neg.)
Multi-task, overlap removed (exact SMILES)	0.490 ± 0.017	−0.087 ± 0.886 [0.397] (8 neg.)
Tuned single-task NN	0.491 ± 0.012	−0.377 ± 1.650 [0.383] (9 neg.)
Tuned multi-task NN	0.497 ± 0.021	−0.012 ± 0.817 [0.399] (9 neg.)
Gradient boosting (reference)	0.581 ± 0.014	0.388 ± 0.188 [0.482] (1 neg.)

Values are mean ± SD; scaffold medians are in brackets. "neg." denotes partitions with negative R^2.

3.5. Neural-Network Instability and Atypical Compounds

To determine whether the neural-network instability reflected a genuine architectural limitation rather than an artifact of the splitting implementation, we examined which scaffold-held-out test folds produced negative R^2. The single worst partition (seed 5, single-task $R^2 = -8.94$) contained in its test fold xenon (CAS 7440-63-3, SMILES [Xe], a single-atom compound with a real, experimentally reported logBB = 0.03 in B3DB, reflecting its clinical use as an anesthetic agent [10]. Xenon's average molecular weight descriptor (AMW = 131.9) lay, under the training-partition RobustScaler fit, about 92 scaled units from the training-set median, a magnitude far outside the range encountered in training. A systematic search of B3DB with an objective drug-likeness filter (fewer than four heavy atoms or no carbon atom) identified 20 such compounds (noble gases of neon, argon, krypton, and xenon; the alkali metal lithium; small gases and volatile anesthetics—methane, ethylene, nitrogen (two B3DB entries, one of which carries a SMILES string that RDKit parses as hydrazine rather than dinitrogen), nitrous oxide (two entries), water, methanol, ethanol, dichloromethane, carbon disulfide, cyclopropane, propene, oxirane, and sulfur hexafluoride; Table 5), all of which are legitimate B3DB entries with real measured logBB values but which lie far outside the physicochemical space of the drug-like organic molecules that dominate the dataset.

Eighteen of the 20 compounds had no ring and therefore shared a single Murcko group (the empty scaffold, 117 acyclic compounds, the largest scaffold group in the dataset); only cyclopropane and oxirane had ring scaffolds. Because the scaffold groups were assigned as units, these 18 compounds entered the test fold. This happened in 13 of the 30 partitions. In these 13 partitions, the largest scaled descriptor deviation in the test fold was 88–96 units

(xenon, AMW), and the single-task network had a mean R^2 of −2.47 (median −1.34; negative in 10 of 13), the multi-task network −0.70 (median −0.74; negative in 10 of 13), and gradient boosting 0.23 (median 0.18; negative in one of 13) (Table S8). In the other 17 partitions, the largest deviation was 18–26 units, and no model produced a negative R^2 (single-task 0.449, multi-task 0.437, gradient boosting 0.511; medians 0.448, 0.442, and 0.528). Across partitions, the R^2 of the single-task network was strongly rank-correlated with the largest scaled deviation (Spearman's $\rho = -0.78$, $p = 5 \times 10^{-7}$). Gradient boosting was also less accurate when the acyclic group was in the test fold (0.23 versus 0.51), so this group is intrinsically harder to predict, but its failures were far smaller than those of the networks, because tree ensembles partition feature space by thresholds and are inherently less sensitive to extreme covariate values than standardized dense networks. Z-score standardization did not behave differently: the single-task network gave −0.619 ± 1.974 (negative in 10 of 30 partitions; minimum −7.25) with Z-score scaling versus −0.817 ± 2.412 (10 of 30; minimum −8.94) with RobustScaler; so, the instability was not an artifact of the scaler choice (Table S9).

Table 5. Non-drug-like and compositionally atypical B3DB compounds associated with neural-network scaffold-split instability.

B3DB Compound Label	SMILES	Heavy Atoms	logBB (Experimental)
Lithium	[Li]	1	−0.30
Krypton	[Kr]	1	−0.16
Methane	C	1	0.00
Water	O	1	−0.04
Neon	[Ne]	1	0.20
Argon	[Ar]	1	0.03
Xenon	[Xe]	1	0.03
Methanol	CO	2	0.02
Ethylene	C=C	2	0.30
Nitrogen	N#N	2	0.03
Nitrogen *	NN	2	0.03
Nitrous oxide, 99%	N#[N+]O	3	0.03
Nitrous oxide	[N−]=[N+]=O	3	0.03
Oxirane	C1CO1	3	0.01
Carbon disulfide	C(=S)=S	3	0.60
Propene	CC=C	3	−0.06
Dichloromethane	C(Cl)Cl	3	−0.30
Ethanol	CCO	3	−0.20
Cyclopropane (trimethylene)	C1CC1	3	0.10
Elegas (sulfur hexafluoride)	FS(F)(F)(F)(F)F	7	0.40

Compound labels are reproduced verbatim from B3DB. * This second "nitrogen" entry carries the SMILES "NN", which RDKit parses as hydrazine (N_2H_4) rather than dinitrogen (N#N); this is a labelling inconsistency in the source database, noted here for transparency rather than correction. Several of these compounds (xenon, nitrous oxide) are volatile anesthetics with genuine pharmacological relevance to CNS penetration and are correctly retained in B3DB; their listing here documents a model-robustness consideration rather than a data-quality concern.

As shown in Figure 5A, the original dataset ($n = 1058$) exhibited substantial neural-network instability under scaffold splitting, with mean R^2 values of −0.817 ± 2.412 for the single-task network and −0.054 ± 0.787 for the multi-task network. After removing the 20 atypical compounds ($n = 1038$), all models were retrained (Tables 6 and S10, Figure 5B). The random-split R^2 was 0.589 ± 0.007 (gradient boosting), 0.505 ± 0.015 (single-task), and 0.514 ± 0.022 (multi-task). Under scaffold splitting, the neural-network instability largely disappeared: the single-task network's mean R^2 increased to 0.346 ± 0.168 (negative partitions 10/30 → 2/30; minimum −0.09), and that of the multi-task network increased to 0.409 ± 0.091 (10/30 → 0/30; minimum 0.16), while gradient boosting achieved 0.420 ±

0.164 (1/30 negative; median 0.465). On the cleaned dataset, the multi-task network was statistically indistinguishable from gradient boosting (difference GB − multi-task = +0.011, 95% CI −0.043 to +0.058; Cohen's dz = 0.08), whereas the advantage of gradient boosting over the single-task network was small (+0.074, 95% CI +0.024 to +0.119; dz = 0.54; better in 23 of 30 partitions; corrected $p = 0.35$). The multi-task network was nominally better than the single-task network (+0.063, 95% CI +0.014 to +0.119; paired t-test $p = 0.030$), but this difference was supported neither by the Wilcoxon signed-rank test ($p = 0.073$) nor by the dependence-corrected test ($p = 0.47$); the corresponding median R^2 values were 0.431 and 0.416, respectively. These findings strongly support an association between the presence of compositionally atypical compounds and neural-network instability. However, because the exclusion criterion was applied after identifying the failure pattern, the cleaned-dataset evaluation should be interpreted as a sensitivity analysis rather than a formal causal test. Furthermore, two scaffold partitions remained negative for the single-task network, indicating that the removal of atypical compounds did not completely eliminate prediction instability.

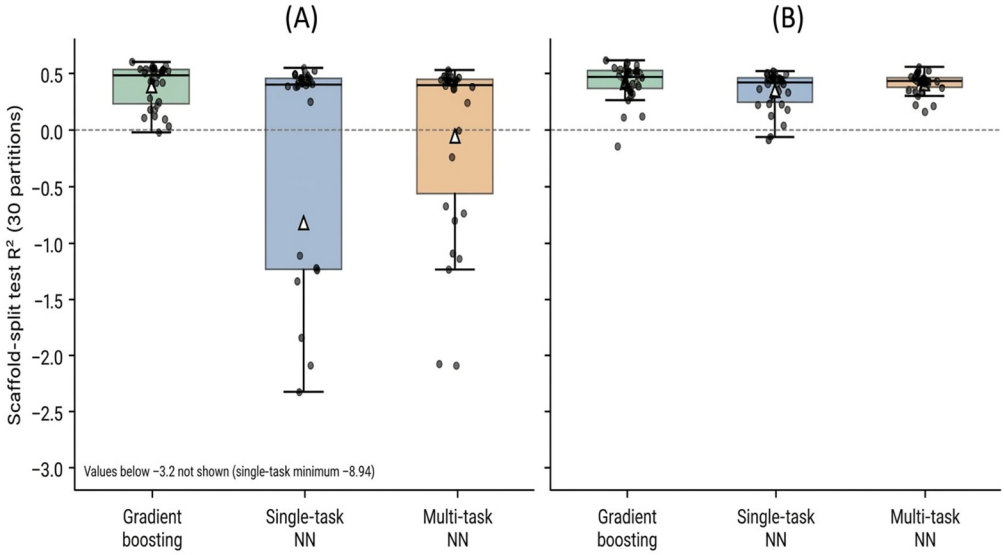

Figure 5. Scaffold-split test R^2 distributions across 30 partitions for gradient boosting, single-task neural network, and multi-task neural network. (**A**) Original B3DB dataset ($n = 1058$). (**B**) Cleaned dataset after excluding 20 atypical compounds ($n = 1038$). Boxes indicate interquartile ranges, horizontal lines indicate medians, triangles represent means, and points represent individual partitions. The dashed line indicates $R^2 = 0$. Values below −3.2 are omitted from panel (**A**) for clarity (minimum single-task $R^2 = -8.94$).

Table 6. Performance on the cleaned dataset (20 atypical compounds removed; $n = 1038$) compared with the full dataset (mean ± SD; negative scaffold partitions in parentheses).

Split	Model	Full Dataset ($n = 1058$)	Cleaned ($n = 1038$)
Random (10 seeds)	Gradient boosting	0.581 ± 0.014	0.589 ± 0.007
	Single-task NN	0.482 ± 0.015	0.505 ± 0.015
	Multi-task NN	0.485 ± 0.016	0.514 ± 0.022
Scaffold (30 partitions)	Gradient boosting	0.388 ± 0.188 (1 neg.)	0.420 ± 0.164 (1 neg.)
	Single-task NN	−0.817 ± 2.412 (10 neg.)	0.346 ± 0.168 (2 neg.)
	Multi-task NN	−0.054 ± 0.787 (10 neg.)	0.409 ± 0.091 (0 neg.)

3.6. Held-Out Classification Results

On the original external set after removing parent-structure duplicates of the regression set ($n = 153$; 151 BBB^+, 2 BBB^-), the ten-seed mean predictions gave AUC-ROC = 0.854 (95%

CI 0.689–0.987) for the single-task network, 0.841 (0.673–0.980) for the multi-task network, and 0.772 (0.530–0.987) for gradient boosting (Figure 6A; Table 7). These intervals overlap almost completely and rest on two negative compounds (four in the full 175-compound set, Table S11); no ranking of models is claimed. On the expanded held-out set (n = 3031 unique parent structures; 1651 BBB^+, 1380 BBB^-), AUC-ROC was 0.764 (95% CI 0.748–0.781) for the single-task network, 0.772 (0.756–0.789) for the multi-task network, and 0.767 (0.750–0.784) for gradient boosting (Figure 6B; Table 7). The paired bootstrap difference between the multi-task network and gradient boosting was +0.005 (95% CI −0.004 to +0.014) and between the single-task network and gradient boosting −0.003 (−0.012 to +0.005); only the small difference between the multi-task and single-task network excluded zero (+0.008, 95% CI +0.005 to +0.011) and is too small to be of practical importance.

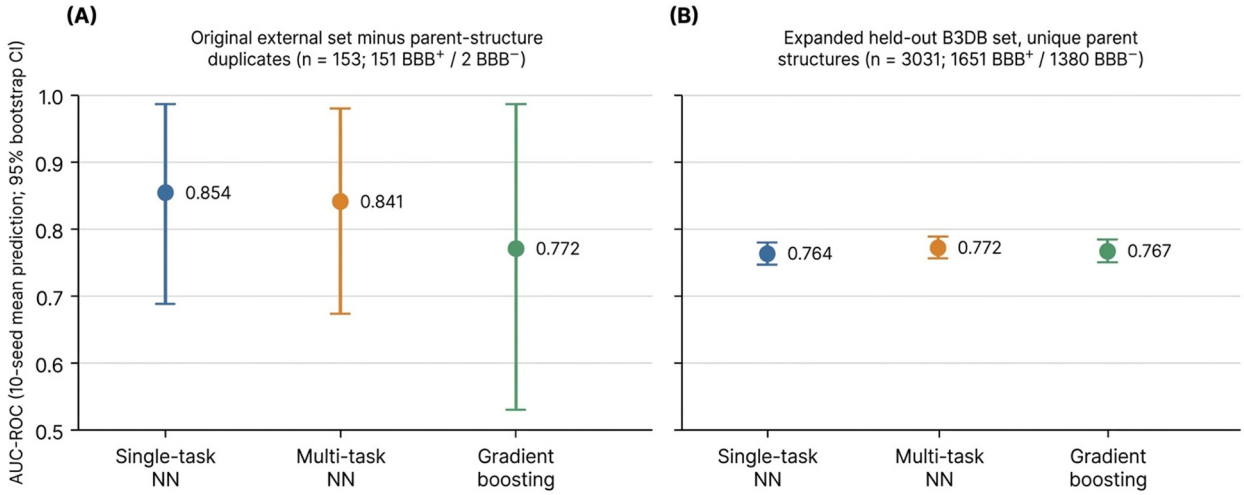

Figure 6. AUC-ROC comparison of single-task neural network, multi-task neural network, and gradient boosting models on held-out BBB classification datasets. (**A**) Original external set after parent-structure deduplication (n = 153; 151 BBB^+/2 BBB^-). (**B**) Expanded held-out B3DB set (n = 3031; 1651 BBB^+/1380 BBB^-). Points represent AUC-ROC from ten-seed averaged predictions; error bars indicate 95% bootstrap confidence intervals (2000 resamples).

Table 7. Held-out classification evaluation (TP/FN refer to BBB^+, TN/FP to BBB^-; ten-seed mean predictions; CIs from 2000 bootstrap resamples of compounds).

Set	Model	AUC-ROC [95% CI]	TP/FN/TN/FP, logBB = 0	Bal. acc.	TP/FN/TN/FP, logBB = −1	Bal. acc.
Original set minus parent duplicates (n = 153; 151+/2−)	Single-task NN	0.854 [0.689, 0.987]	85/66/2/0	0.781	147/4/0/2	0.487
	Multi-task NN	0.841 [0.673, 0.980]	92/59/2/0	0.805	146/5/0/2	0.483
	Gradient boosting	0.772 [0.530, 0.987]	80/71/2/0	0.765	145/6/1/1	0.730
Expanded, unique parent structures (n = 3031; 1651+/1380−)	Single-task NN	0.764 [0.748, 0.781]	799/852/1120/260	0.648	1623/28/231/1149	0.575
	Multi-task NN	0.772 [0.756, 0.789]	858/793/1102/278	0.659	1626/25/235/1145	0.578
	Gradient boosting	0.767 [0.750, 0.784]	723/928/1161/219	0.640	1605/46/340/1040	0.609

TP, true positive; FN, false negative; TN, true negative; FP, false positive; Bal. acc., balanced accuracy.

Results were similar for compounds whose parent structure is absent from the hERG training set (n = 1984: AUC 0.796, 0.806, and 0.802) and for those present in it (n = 1047:0.723, 0.731, and 0.724). Threshold-dependent performance depended strongly on the cut-off: at logBB = −1, balanced accuracy was 0.575 (single-task), 0.578 (multi-task), and 0.609 (gradient boosting), because gradient boosting identified more BBB^- compounds (specificity 0.246 versus 0.167 and 0.170) at similar sensitivity (0.972–0.985); at logBB = 0

balanced accuracy was 0.648, 0.659, and 0.640. Because 107 of the 3324 parent structures in the unfiltered classification set carried conflicting labels among their records, and labels are not homogeneous, these results should be read as an expanded held-out check and not as definitive external validation. Collapsing duplicates was important: before parent-structure collapsing (6771 records), the AUCs were 0.827, 0.840, and 0.828, and the multi-task advantage over gradient boosting appeared larger (+0.012), an effect attributable to repeated stereoisomers and salts of the same compounds.

3.7. SHAP-Based Model Interpretation

The SHAP analysis of the Bayesian-optimized gradient-boosting model on the 212 held-out random-split test compounds identified topological polar surface area (TopoPSA) as the dominant predictor of experimental logBB (mean |SHAP| = 0.344 log units; Table 8, Figure 7), about 3.5 times the contribution of the second-ranked feature, SLogP (0.097), followed by double-bond count (nBondsD, 0.061), acidic group count (nAcid, 0.045), and basic group count (nBase, 0.041), reproducing the previously reported top-five feature set [2]. Permutation importance confirmed TopoPSA, SLogP, nAcid, and nBase as the top four features (nBondsD ranked seventh; Tables 8 and S12). These findings are consistent with established structure–permeability relationships [2,7]. The analysis serves only to show that Bayesian re-optimization did not alter the model's mechanistic feature hierarchy.

Table 8. SHAP and permutation feature importance, top 10 features (Bayesian-optimized gradient boosting, test set $n = 212$).

Rank	Feature (SHAP)	Mean |SHAP| (log Units)	Feature (Permutation)	Permutation Importance (ΔR^2)
1	TopoPSA	0.344	TopoPSA	0.543
2	SLogP	0.097	SLogP	0.077
3	nBondsD	0.061	nAcid	0.049
4	nAcid	0.045	nBase	0.033
5	nBase	0.041	BalabanJ	0.025
6	PEOE_VSA1	0.027	SlogP_VSA3	0.021
7	BertzCT	0.026	nBondsD	0.021
8	PEOE_VSA2	0.026	PEOE_VSA2	0.016
9	SlogP_VSA3	0.026	Kier3	0.016
10	BalabanJ	0.026	SlogP_VSA1	0.015

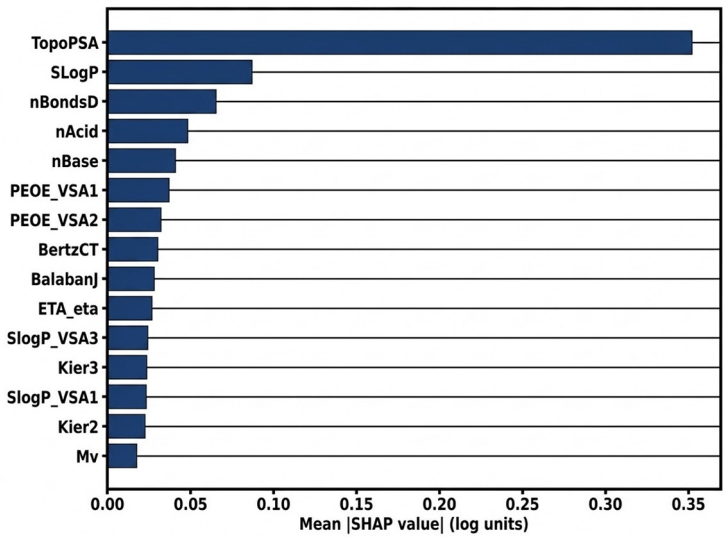

Figure 7. SHAP global feature importance (top 15 features), Bayesian-optimized gradient boosting model, random-split held-out test set ($n = 212$).

4. Discussion

The central motivating hypothesis of this study, that auxiliary supervision from a mechanistically related, larger endpoint (hERG inhibition) would improve BBB permeability prediction via a shared representation, was not supported for logBB regression. This is informative because the benefits of multi-task learning are known from the broader ADMET literature to be inconsistent and task-pair-dependent [5,14]. hERG is moderately related to logBB ($\rho = 0.32$ on shared compounds), but the relationship runs through lipophilicity and basicity rather than polarity, the dominant determinant of logBB (TopoPSA); this may explain why a shared encoder gains little. The null result was robust to loss weighting, neural-network tuning, and removal of overlapping compounds (Section 3.4). The only consistent effect was a reduction in the severity of scaffold-split failures by the multi-task network (Section 3.2), which was not accompanied by better typical performance and did not appear in the cleaned dataset as a robust difference (Section 3.5). On the expanded held-out classification set, the multi-task network was marginally better than the single-task network (ΔAUC +0.008) and not different from gradient boosting (Section 3.6). Alternative auxiliary endpoints more tightly coupled to BBB permeability (plasma protein binding, P-glycoprotein efflux), adaptive loss weighting, and curriculum training remain open and are not excluded by these results.

4.1. Hyperparameter Optimization and Scaffold-Split Robustness

Bayesian optimization selected a deep gradient-boosting configuration (249 estimators, maximum depth 8) that matched the simpler configuration under random splitting (0.581 versus 0.582) but was somewhat less robust under scaffold splitting (mean R^2 0.388 versus 0.440; Wilcoxon $p = 0.34$ and corrected $p = 0.49$; so, the difference is suggestive rather than established). A plausible explanation is that deeper trees have greater capacity to fit fine-grained patterns specific to the scaffold composition of the training data, which may benefit interpolation but be detrimental when extrapolating to novel scaffolds. These findings suggest that hyperparameter selection based solely on random-split or conventional cross-validation performance may not ensure robustness under structural distribution shift, and that scaffold-split performance should be considered during model selection when prospective scaffold-hopping is the intended application. The comparison with the neural networks is between a Bayesian-optimized gradient-boosting model and networks with a fixed or only lightly tuned configuration (Section 3.4); the observed differences should therefore not be read as an intrinsic advantage of either model family.

4.2. Neural-Network Robustness Under Scaffold Splitting

A standardized dense feed-forward neural network using the same 2D descriptor set that performs well with gradient boosting showed substantial instability under scaffold-split evaluation: across 30 partitions, the networks had more frequent and far larger negative R^2 values than gradient boosting, although the corrected test did not establish the mean difference (Section 3.2). The main contributor identified here was a group of 18 acyclic, compositionally atypical compounds that share one scaffold group and therefore enter the test fold together (Section 3.5). Whenever this group was in the test fold, the networks failed (negative R^2 in 10 of 13 partitions), whereas in all other partitions, no model failed; removing the compounds brought the multi-task network to the level of gradient boosting (Section 3.5). This supports a strong association between these compounds and the failures; it is a sensitivity analysis and not a proof of causation, and gradient boosting itself was less accurate on this scaffold group. The original external set contained none of these compounds (minimum heavy-atom count 7), and the expanded held-out set only 3 of 3031 (0.1%) under the same structural rule, while neither showed a neural-network deficit (AUC

0.764–0.772 versus 0.767; Section 3.6); these sets are therefore consistent with the extrapolation explanation but do not test it independently. We therefore interpret the instability as a targeted extrapolation failure, best handled by applicability-domain screening [21] or by excluding non-drug-like compounds from model development, rather than as a general weakness of neural networks. A practical corollary is that the scaffold group of acyclic compounds should be inspected before scaffold-split results are interpreted.

4.3. Structural Overlap and Multi-Task Evaluation

Between 30% and 49% of the BBB test compounds shared a parent structure with the hERG training set (39.6% on the random split); thus, the shared encoder had indirect exposure to the structures of a large fraction of nominally held-out compounds. This does not constitute label leakage because hERG activity and logBB are independently measured endpoints. Although structural overlap could favor beneficial transfer, no multi-task advantage was observed (Sections 3.1 and 3.2), and retraining with all overlapping compounds removed from the hERG training set gave the same result (Section 3.4). Therefore, the overlap does not explain the null result, but it shows that scaffold-split test sets are not independent of an auxiliary dataset unless the two are deduplicated, which we recommend for future multi-task studies. Matching by exact SMILES understated the overlap (29.7% versus 39.6%) because salts and stereoisomers are not recognized as having the same structure.

4.4. Implications for BBB Modelling and Future Directions

The present study provides a systematic evaluation of multi-task learning for blood–brain barrier (BBB) permeability prediction using hERG inhibition as an auxiliary endpoint. By integrating repeated random- and scaffold-split evaluations, statistical comparisons, overlap-removal experiments, and sensitivity analyses, the study highlights the importance of model robustness under structural distribution shifts. Consistent with recent advances in graph-based BBB prediction and curated molecular benchmarks [10–12], the findings emphasize the need for chemically informed validation. For descriptor-based models trained on approximately 1000 compounds, gradient boosting remained a competitive and relatively stable baseline [13], whereas hERG-assisted multi-task learning provided no consistent improvement in continuous logBB prediction or held-out classification performance.

Although additional loss-weight experiments and hyperparameter searches supported the main conclusions, future studies could investigate adaptive task weighting, nested scaffold-aware optimization, and graph-based molecular representations. Because the repeated scaffold partitions originated from the same dataset, dependence-corrected statistical comparisons should be interpreted cautiously. An important finding was the association between compositionally atypical compounds and neural-network instability. Their removal substantially improved scaffold-split performance, particularly for the multi-task model; however, this post hoc analysis warrants prospective validation using applicability-domain screening, out-of-distribution detection, and uncertainty quantification.

The held-out classification results further highlight the need for larger, independently measured, class-balanced datasets with standardized BBB endpoints. Moreover, hERG inhibition represents only one aspect of pharmaceutical safety, while BBB exposure and cardiovascular responses involve additional mechanisms, including P-glycoprotein-mediated efflux, plasma protein binding, and mitochondrial quality-control pathways associated with myocardial ischemia–reperfusion injury [22]. Future multi-endpoint frameworks incorporating complementary pharmacokinetic and toxicity endpoints may help clarify when auxiliary-task learning benefits BBB prediction. Alternative descriptor sets and molecular representations also warrant investigation, as the present study retained the previously pub-

lished molecular descriptors [2]. Overall, these findings establish the importance of dataset composition, scaffold-aware validation, and model robustness when assessing multi-task learning strategies for pharmaceutical property prediction.

5. Conclusions

The principal conclusions of this study are as follows:

1. Multi-task learning with hERG did not improve logBB regression. Under random splitting, the multi-task and single-task networks were indistinguishable ($R^2 = 0.485$ versus 0.482, $p = 0.69$). Under scaffold splitting, the multi-task network had a higher mean R^2 only through less severe failures (negative R^2 in 10 of 30 partitions for both networks), with nearly identical medians (0.394 versus 0.398). No loss weight, tuning setting, or removal of 30–49% of test compounds sharing a parent structure with hERG training data revealed a multi-task benefit, and AUCs on 3031 held-out parent structures were similar (0.764–0.772).
2. Random-split optimization did not guarantee scaffold-split robustness: the Bayesian-optimized gradient-boosting model matched a simpler configuration on random splits ($R^2 = 0.581$ versus 0.582) but was somewhat less robust across scaffold splits (0.388 versus 0.440).
3. Gradient boosting was more robust than the neural networks under scaffold splitting in 28–29 of 30 partitions (mean ΔR^2 0.44–1.21), a consistent effect that was not significant under a dependence-corrected test and was concentrated in the lower tail of the neural network results.
4. Neural network instability was strongly associated with a scaffold group of acyclic, compositionally atypical compounds: networks failed in 10 of the 13 partitions in which this group was held out and in none of the others, and after removing 20 such compounds, the multi-task network ($R^2 = 0.409$) matched gradient boosting (0.420). This was a sensitivity analysis, not a formal causal test.
5. Scaffold-split evaluation, applicability-domain screening, and deduplication between auxiliary and target data should be incorporated into the BBB model development; at the present scale, curated 2D-descriptor gradient boosting provides a robust baseline, and neural networks should be stress-tested for outlier-driven failures before use in scaffold-hopping virtual screening.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pharmaceutics18111449/s1. Table S1: Compound-level predictions; Table S2: Per-run model performance; Table S3: Structural overlap analysis; Tables S4 and S5: Overlap-removal analyses; Table S6: Loss-weight sensitivity; Table S7: Neural-network optimization; Table S8: Atypical-compound partition analysis; Table S9: Z-score standardization; Table S10: Cleaned-dataset performance; Table S11: Held-out classification results; Table S12: SHAP and permutation importance.

Author Contributions: Conceptualization, S.V.P.V. and S.T.; methodology, S.V.P.V. and S.T.; software, S.T., S.V.P.V. and P.S.; validation, S.V.P.V. and P.S.; formal analysis, S.V.P.V., S.T. and P.S.; investigation, P.S., S.V.P.V. and S.T.; resources, S.T. and S.V.P.V.; data curation, P.S. and S.T.; writing—original draft preparation, S.V.P.V. and S.T.; writing—review and editing, S.T. and S.V.P.V.; visualization, P.S. and S.V.P.V.; supervision, S.T. and S.V.P.V.; project administration, S.T. and S.V.P.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The B3DB and hERG datasets analyzed in this study are publicly available from the repositories cited in the manuscript. Compound-level experimental and predicted logBB values for the held-out random-split test set are provided in Supplementary Tables.

Further information and supporting data are available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jiang, Y.; Lam, S.M.; Zhang, S.; Miao, H.; Zhou, Y.; Zhang, Q.; Zhou, T.; Feng, H.; Ding, N.; Wang, H.; et al. CSF Multi-Omics of Intracerebral Hemorrhage from Onset to Reperfusion Underscores Lipid Metabolism in Functional Outcome. *Sci. Bull.* **2025**, *70*, 162–166. [CrossRef] [PubMed]
2. Tiwari, S.; Mądra-Gackowska, K.; Gackowski, M.; Park, N.; Szeleszczuk, Ł. Predicting Blood–Brain Barrier Permeability from Experimental Data: An Interpretable and Externally Validated Machine Learning Framework. *Pharmaceutics* **2026**, *18*, 670. [CrossRef] [PubMed]
3. Tiwari, S.; Mądra-Gackowska, K.; Gackowski, M.; Szeleszczuk, Ł. Quantitative Structural Thresholds for Blood–Brain Barrier Permeability Derived from Experimental LogBB Measurements. *Pharmaceutics* **2026**, *18*, 967. [CrossRef] [PubMed]
4. Tiwari, S.; Mądra-Gackowska, K.; Gackowski, M.; Szeleszczuk, Ł. Evaluating the Incremental Value of Three-Dimensional Conformer Descriptors for Blood–Brain Barrier Permeability Prediction Using the B3DB Benchmark Dataset. *Pharmaceutics* **2026**, *18*, 1018. [CrossRef] [PubMed]
5. Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
6. Sanguinetti, M.C.; Tristani-Firouzi, M. HERG Potassium Channels and Cardiac Arrhythmia. *Nature* **2006**, *440*, 463–469. [CrossRef] [PubMed]
7. Hitchcock, S.A.; Pennington, L.D. Structure−Brain Exposure Relationships. *J. Med. Chem.* **2006**, *49*, 7559–7583. [CrossRef] [PubMed]
8. Li, W.-Q.; Tan, S.-L.; Li, X.-H.; Sun, T.-L.; Li, D.; Du, J.; Wei, S.-S.; Li, Y.-J.; Zhang, B.-K. Calcitonin Gene-Related Peptide Inhibits the Cardiac Fibroblasts Senescence in Cardiac Fibrosis via up-Regulating Klotho Expression. *Eur. J. Pharmacol.* **2019**, *843*, 96–103. [CrossRef] [PubMed]
9. Jiang, C.; Xie, N.; Sun, T.; Ma, W.; Zhang, B.; Li, W. Xanthohumol Inhibits TGF-B1-Induced Cardiac Fibroblasts Activation via Mediating PTEN/Akt/MTOR Signaling Pathway. *Drug Des. Devel. Ther.* **2020**, *14*, 5431–5439. [CrossRef] [PubMed]
10. Meng, F.; Xi, Y.; Huang, J.; Ayers, P.W. A Curated Diverse Molecular Database of Blood-Brain Barrier Permeability with Chemical Descriptors. *Sci. Data* **2021**, *8*, 289. [CrossRef] [PubMed]
11. Kumar, S.; Yadav, S.; Vora, D.S.; Paul, P.; Dhanjal, J.K. GraphB3—An Interpretable Graph Learning Approach for Predicting Blood–Brain Barrier Permeability. *Brief. Bioinform.* **2025**, *26*, bbaf679. [CrossRef] [PubMed]
12. Ding, Y.; Jiang, X.; Kim, Y. Relational Graph Convolutional Networks for Predicting Blood–Brain Barrier Penetration of Drug Molecules. *Bioinformatics* **2022**, *38*, 2826–2831. [CrossRef] [PubMed]
13. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? In *Advances in Neural Information Processing Systems 35*; NeurIPS: San Diego, CA, USA, 2022.
14. Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 1253–1268. [CrossRef] [PubMed]
15. Siramshetty, V.B.; Nguyen, D.-T.; Martinez, N.J.; Southall, N.T.; Simeonov, A.; Zakharov, A.V. Critical Assessment of Artificial Intelligence Methods for Prediction of HERG Channel Inhibition in the "Big Data" Era. *J. Chem. Inf. Model.* **2020**, *60*, 6007–6019. [CrossRef] [PubMed]
16. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **2018**, *10*, 4. [CrossRef] [PubMed]
17. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
18. Bemis, G.W.; Murcko, M.A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. [CrossRef] [PubMed]
19. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U., Von, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
20. Nadeau, C.; Bengio, Y. Inference for the Generalization Error. *Mach. Learn.* **2003**, *52*, 239–281. [CrossRef]

21. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810. [CrossRef] [PubMed]
22. Deng, J.; Liu, Q.; Ye, L.; Wang, S.; Song, Z.; Zhu, M.; Qiang, F.; Zhou, Y.; Guo, Z.; Zhang, W.; et al. The Janus Face of Mitophagy in Myocardial Ischemia/Reperfusion Injury and Recovery. *Biomed. Pharmacother.* **2024**, *173*, 116337. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.