Discovery report for admet2

Research Objective

Develop an excellent ML model for predicting the following properties:
LogD: Measures compound's lipophilicity at a specific pH. Drugs typically fall into a well defined LogD range that balances aqueous solubility with membrane permeability, making understanding changes in LogD across a chemical series important to medicinal chemists. Additionally, assessing LogD of candidate molecules can suggest whether candidate molecules are "efficient" for their lipophilicity (lipophilicity generally has a linear relationship with affinity). Kinetic Solubility (KSOL): Measures how much a compound can be dissolved under non-equilibrium conditions. Helps screen compounds that will fail due to poor absorption or low bioavailability. Human Liver Microsomal (HLM) stability: Supports understanding of a compound's susceptibility to liver metabolism and can be used to predict in-vivo clearance of a candidate molecule. Measured using human liver microsomes and reported as intrinsic clearance for the compound Clint (mL/min/kg). Mouse Liver Microsomal (MLM) stability: Also used to predict in-vivo clearance of a candidate. The study of both MLM and RLM can provide a more comprehensive understanding of a compound's metabolic profile and how a compound might behave in multi-species preclinical development. Caco-2 Papp A>B: Measures the rate of flux of a compound across polarized Caco-2 cell monolayers from the apical (intestinal lumen-facing side) to basolateral (blood-facing side). This effectively mimics the absorption of a drug across the intestinal wall. Caco-2 Efflux Ratio: Measures the rate of flux of a compound across polarized Caco-2 cell monolayers. The Efflux Ratio is determined by a ratio of the apparent permeability coefficient (Papp) in both directions. Ratios of ~1 indicate that a compound primarily traverses the cell membrane via passive (diffusional) transport. Ratios > 2 generally indicate active transport of the compound across the cellular membrane by membrane bound transporters (e.g. efflux by p-glycoprotein). Mouse Plasma Protein Binding (MPPB): Determines the concentration of free drug in plasma (as % Unbound). Drugs that are not bound to plasma can bind to target proteins and yield the desired therapeutic effect, making this parameter crucial to understanding drug distribution. Mouse Brain Protein Binding (MBPB):
This measures the fraction of a drug not bound to proteins within brain tissue. The unbound fraction of a drug in the brain is considered pharmacologically active and able to interact with central nervous system (CNS) targets. MBPB helps assess CNS drug exposure and potential efficacy or side effects for neuroactive compounds. Reported as % Unbound.
Mouse Gastrocnemius Muscle Binding (MGMB):. This reflects the amount of drug free to act within skeletal muscle tissue, which is important for drugs targeting peripheral or muscular conditions (very important for the DM1 indication). Reported as % Unbound
While training the model, pay close attention to timing information. You are running in a virtual machine with ~32GB of RAM and ~8 CPUs and 1200 seconds per cell timeouts.
You must have predictions in a csv (same format as training) for the test compounds. As soon as possible (after a few iterations), start creating predictions files from the best current models.
On the leaderboard, the best scores are 0.77 spearman/0.59 MA-RAE. You should do a literature search (after dataset exploration, because chemistry tasks require some specific). For example, chemprop is getting good scores. Although be careful about compute time!

Dataset Description

The training dataset and blind test data for your evaluations

Summary of Discoveries

Discovery 1: Data pathology characterization and a robust, minimal preprocessing pipeline

The ADMET dataset shows severe label sparsity, target skewness, and descriptor pathologies that induce model instability and bias. A minimal preprocessing pipeline—replace invalids, standardize, hard-clip, and remove a single toxic descriptor (Ipc)—restores numerical stability and improves accuracy, whereas aggressive feature removal and non-linear rank transforms degrade performance.

Discovery 2: Multi-task learning under label sparsity: benefits, stability, and limits

On a pharmaceutically relevant ADMET panel with severe and structured label missingness, a shared multi-task feed-forward neural network outperformed strong single-task baselines by exploiting cross-endpoint structure. The gains were largest on the sparsest endpoints, but achieving stable training required targeted input standardization, clipping, and removal of a pathological descriptor; explicit hierarchical modeling and complexity-based partitioning did not add value beyond implicit multi-task sharing.

Discovery 3: Heterogeneous, per-target weighted ensembling delivers state-of-the-art performance

Combining a multi-task neural network with single-task boosted trees and blending them with per-target weights yields the highest overall accuracy on a challenging ADMET panel. A LightGBM-based hybrid with a KSOL-specific subsystem and selective target transforms reached a mean validation Spearman correlation of 0.8558, and a GBR fallback maintained robust performance and produced complete test submissions under software constraints.

Discovery 4: Decoupling rank accuracy from calibration: transforms, stratification, and post-hoc calibration

Across nine ADMET endpoints with heavy missingness and skewed distributions, high-performing rank-based models were found to be poorly calibrated in magnitude. Target-aware transformations and post-hoc isotonic regression corrected these magnitude errors without harming rank order, while value-stratified models improved performance selectively when supported by reliable regime classifiers and rich features.

Data pathology characterization and a robust, minimal preprocessing pipeline

Summary

The ADMET dataset shows severe label sparsity, target skewness, and descriptor pathologies that induce model instability and bias. A minimal preprocessing pipeline—replace invalids, standardize, hard-clip, and remove a single toxic descriptor (Ipc)—restores numerical stability and improves accuracy, whereas aggressive feature removal and non-linear rank transforms degrade performance.

Background

Predicting ADMET properties from structure-derived descriptors is central to modern medicinal chemistry, yet real-world datasets often violate modeling assumptions through missing labels, heavy-tailed targets, and distribution shift between discovery and validation chemotypes. Multi-task learning can leverage cross-endpoint correlations to share signal across sparsely labeled tasks, but success hinges on preprocessing that preserves distributed chemical signal while neutralizing pathological features. Robust evaluation using rank-based metrics and carefully defined error measures is essential when targets are non-normal and span multiple orders of magnitude.

Results & Discussion

The dataset comprises 5,326 training and 2,282 test molecules with nine ADMET endpoints and 35.47% missingness overall, including extreme sparsity for MGMB (95.83%), MBPB (81.69%), and MPPB (75.55%) coupled with non-normal, highly skewed targets (Shapiro–Wilk $p < 0.05$; skewness -0.70 to 6.99) and strong cross-endpoint correlations (e.g., MBPB–MGMB $r=0.904$, LogD–MPPB $r=-0.686$) that motivate multi-task learning [r0]. The test set shows clear distribution shift, with longer SMILES strings (57.84 vs 48.03 characters), and the presence of extreme assay outliers (five MLM CLint values $>10{,}000$) further complicates training and evaluation [r0]. Critically, descriptor analysis revealed that 99.56% of test molecules contain at least one standardized feature with $|z| > 10$ across 1,201 of 2,265 features, driven most severely by the Ipc descriptor whose test-set maximum reaches 3.3×10^{18} (standardized to 4.67×10^{9}), which without control induces numerically catastrophic predictions; clipping standardized inputs to $[-10, 10]$ removes 18,688 extreme values in test and 24,094 in train and is essential to recover stable, biologically plausible outputs [r10].

A minimal, principled pipeline emerged: replace inf/NaNs, standardize on the training set, hard-clip to $[-10, 10]$, and remove Ipc only. This intervention stabilized a multi-task feed-forward neural network trained with masked loss on the sparse labels, yielding smooth convergence (final training loss 0.1386) and test-time predictions that are largely within training ranges (e.g., 99.9% for LogD; 98.6% for HLM CLint; 96.4% for MLM CLint), in stark contrast to unclipped models that produced values up to 10^9 in magnitude [r10]. Directly testing the impact of Ipc removal increased mean validation Spearman correlation from 0.7713 to 0.8032 (+0.0319), with 8 of 9 targets improving—including a substantial gain for HLM CLint (+0.1093) and MGMB (+0.0604)—and exhibited more stable optimization dynamics (early stopping at epoch 29 vs 13), confirming Ipc as uniquely toxic to generalization even under clipping [r22].

In contrast, broad feature excision is counterproductive. Removing all 52 descriptors flagged as unstable yielded no benefit over the Ipc-only configuration (mean Spearman 0.8017 vs 0.8032; $\Delta\rho = -0.0015$), indicating that the remaining 51 descriptors retain net signal once standardized and clipped [r27]. Even more aggressively, eliminating 65.74% of features by variance filtering (<0.01) degraded LightGBM performance (mean Spearman 0.7896→0.7596) and dramatically inflated MA-RAE (0.4048→1.9731), with 6 of 9 endpoints worse—consistent with rare but informative substructures being encoded in low-variance fingerprint bits and with boosting's intrinsic robustness to high dimensionality [r7]. Here MA-RAE was computed after excluding

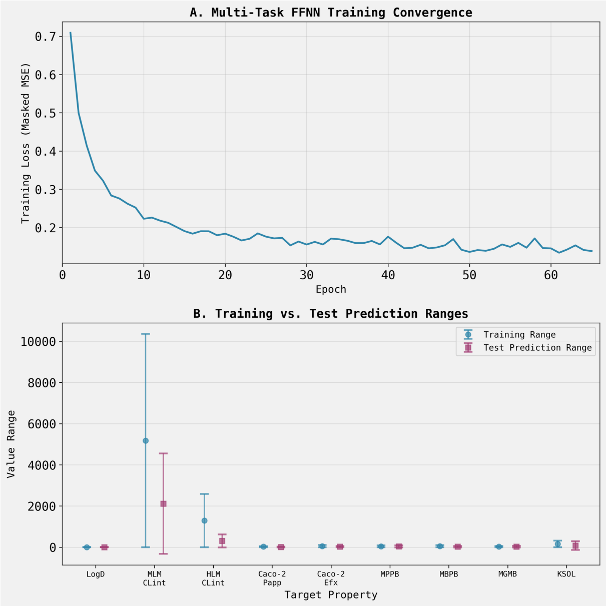


Figure 1: The minimal preprocessing pipeline enables stable model training and controlled prediction outputs. (A) The training loss (masked MSE) of the multi-task feed-forward neural network shows smooth convergence over approximately 65 epochs. (B) A comparison of the predicted value ranges for the training and test sets across the nine ADMET endpoints. The combination of smooth loss decay and bounded, non-catastrophic predictions demonstrates that the pipeline successfully restores numerical stability. (Source: [r10])

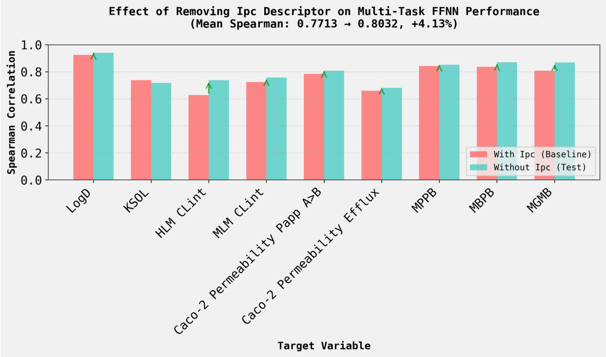


Figure 2: Removing the single, pathologically-scaled Ipc descriptor substantially improves the performance of the multi-task neural network. The bar chart compares the test set Spearman correlation for nine ADMET prediction tasks for a model trained with the full feature set (With Ipc) versus a model trained after removing only the Ipc descriptor (Without Ipc). This single-feature removal increases the mean Spearman correlation from 0.7713 to 0.8032, validating this intervention as a critical component of the minimal preprocessing pipeline for restoring model performance. (Source: [r22])

targets with $|y|$ 0.1 to avoid division by near-zero denominators, aligning error measurement with the properties' heavy-tailed scales [r7].

Alternative non-linear transforms also underperform. Replacing standardization plus clipping with QuantileTransformer led to a sharp drop in mean validation Spearman (0.5443 vs 0.7985), with especially severe losses for Caco-2 permeability (-0.58 and -0.52), and poorer optimization (final loss 0.632 vs 0.088) [r28]. The mechanism is clear: mapping sparse, fingerprint-dominated features (92.47% zeros) to a normal target distribution pushes the modal zero mass into the left tail (mean -4.68), violating the zero-centered input assumption of Xavier-initialized networks and impeding gradient flow; by contrast, standardization plus clipping preserves centering while neutralizing only extreme values [r28]. Training reliability further improved after correcting a masked MSE implementation bug by computing loss on boolean-indexed valid targets, preventing non-differentiable paths and gradient explosions [r28].

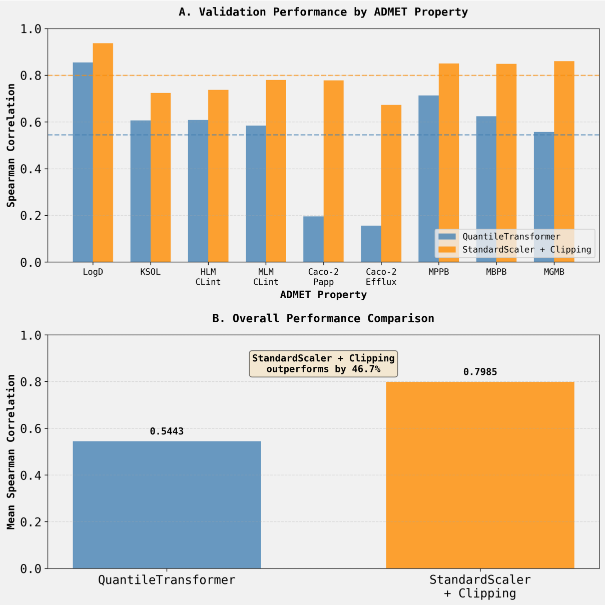


Figure 3: A minimal preprocessing pipeline of standard scaling with clipping substantially outperforms a non-linear quantile transformation. (A) Validation performance, measured by Spearman correlation, is shown for a multi-task model on nine individual ADMET properties after applying either a QuantileTransformer or StandardScaler with clipping to the input features. (B) The mean Spearman correlation across all properties demonstrates a 46.7% improvement for the scaling and clipping approach. These results highlight that a simple, robust preprocessing strategy is superior to aggressive, rank-based transformations for this dataset. (Source: [r28])

Taken together, these results show that the dataset's label sparsity, skewness, and test-time shift interact with descriptor pathologies to create failure modes that are best addressed by a minimal, targeted preprocessing strategy rather than wholesale feature removal or aggressive non-linear transformations. The combination of Ipc removal, standardization, and hard clipping restores numerical stability without sacrificing distributed chemical signal, enabling multi-task models to leverage cross-endpoint structure (e.g., MBPB–MGMB r=0.904) and produce plausible test-time predictions across all nine ADMET tasks despite severe sparsity and shift [r0, r10, r22, r27, r28].

Trajectory Sources

Trajectory r0: This pharmaceutical dataset contains 5,326 training molecules and 2,282 test molecules, with 9 ADMET target variables exhibiting extensive missing data (35.47% overall) and highly skewed distributions requiring specialized modeling approaches.

Trajectory r7: Removing low-variance features (variance < 0.01) from the training dataset resulted in a 3.80% decrease in mean Spearman correlation across all nine ADMET prediction tasks, with 6 of 9 targets showing degraded performance, indicating that variance-based feature selection at this threshold does not i...

Trajectory r10: A multi-task feed-forward neural network (512→256→128 hidden layers, 30% dropout) was successfully trained on 5,326 molecules with clipped standardized features (|z| 10) to predict 9 ADMET properties, achieving a final training loss of 0.1386 and generating biologically plausible predictions for 2...

Trajectory r22: Removing the Ipc descriptor from the feature set improved multi-task FFNN performance, increasing mean Spearman correlation from 0.7713 to 0.8032 (+0.0319, +4.13% relative improvement) on the validation set.

Trajectory r27: Removing all 52 unstable descriptors does not improve multi-task FFNN performance compared to removing only the Ipc descriptor, with mean Spearman correlations of 0.8017 versus 0.8032, respectively ($\Delta\rho$ = -0.0015).

Trajectory r28: QuantileTransformer significantly underperforms compared to StandardScaler + clipping for multi-task ADMET prediction, achieving a mean validation Spearman correlation of 0.5443 versus 0.7985 (46.7% relative decrease in performance).

Multi-task learning under label sparsity: benefits, stability, and limits

Summary

On a pharmaceutically relevant ADMET panel with severe and structured label missingness, a shared multi-task feed-forward neural network outperformed strong single-task baselines by exploiting cross-endpoint structure. The gains were largest on the sparsest endpoints, but achieving stable training required targeted input standardization, clipping, and removal of a pathological descriptor; explicit hierarchical modeling and complexity-based partitioning did not add value beyond implicit multi-task sharing.

Background

ADMET properties such as lipophilicity, solubility, permeability, metabolic stability, and tissue protein binding jointly determine exposure, distribution, and ultimately the translational viability of small molecules. In modern medicinal chemistry campaigns, high-throughput profiling produces partially observed matrices of endpoints across chemical series, creating a natural setting for multi-task learning to leverage shared structure while contending with label sparsity and skewed distributions. Recent graph and message-passing architectures advocate shared encoders with task-specific heads, descriptor fusion, and loss masking for missing labels; however, compute-efficient dense models trained on engineered descriptors remain practical alternatives when datasets are modest in size and label missingness is extreme.

Results & Discussion

The dataset comprised 5,326 training molecules and 2,282 test molecules with nine ADMET targets and extensive missingness (35.47% overall), with endpoint-specific gaps ranging from 5.39% (LogD) to 95.83% (Mouse Gastrocnemius Muscle Binding, MGMB; 222 samples) and only 2.33% of compounds fully profiled across all endpoints [r0]. Distributions were strongly non-normal (Shapiro–Wilk $p < 0.05$) and skewed, and several endpoint pairs exhibited substantial correlation consistent with pharmacological expectations, including LogD–MPPB ($r = -0.686$), LogD–KSOL ($r = -0.542$), LogD–MBPB ($r = -0.507$), MBPB–MGMB ($r = 0.904$), MPPB–MBPB ($r = 0.614$), and HLM–MLM CLint ($r = 0.561$) [r0]. The blind test set showed a shift toward more complex structures (longer SMILES; 57.84 vs 48.03 characters), underscoring the need for robust generalization within a potentially shifted yet related chemical space [r0].

A comprehensive featurization pipeline yielded 217 RDKit 2D descriptors and 2,048 radius-2 Morgan fingerprint bits per molecule (2,265 features total), with 100% SMILES validity and expected fingerprint sparsity (97.45% zeros; ~52 bits set per molecule) [r1]. Despite this, a sizeable fraction of features were low-variance (65.7%), and the Ipc descriptor exhibited extreme dispersion (variance 5.08×10ˆ17), both suggesting risks for instability and overfitting without careful preprocessing [r1]. As a strong single-task baseline, independent LightGBM models trained per endpoint achieved a mean Spearman correlation of 0.7785 (median 0.8020) across endpoints; performance was excellent for LogD ($R = 0.9467$), MPPB (0.8389), MGMB (0.8371), MBPB (0.8151), and Caco-2 Papp A>B (0.8020), and moderate for Caco-2 Efflux (0.6411) and HLM CLint (0.6318). The accompanying error metric, MA-RAE (lower is better), averaged 1.5851 (median 1.1566), with elevated error for HLM CLint (4.7123) reflecting biological complexity [r2]. These results established descriptor-based learning as competitive and computationally efficient within this setting [r2].

A multi-task feed-forward neural network (shared core with task-specific heads), trained with masked loss to exploit all samples despite missing labels, improved the mean Spearman to 0.8133, a +0.0348 absolute (+4.47% relative) gain over the LightGBM baseline, with all endpoints above $R = 0.71$ and the largest improvements on the sparsest targets (MPPB 0.8662, MBPB 0.8571, MGMB 0.9142; all $p < 1 \times 10$ˆ-17) [r9]. Stabilization was essential: features were standardized and clipped to $[-10, 10]$ to control extreme outliers—particularly the Ipc descriptor—which otherwise produced implausible predictions; with clipping, the trained model generated biologically plausible blind predictions that largely fell within training ranges and were released as ffnn$_{\text{multitask}}$_predictions.csv [r10]. In a targeted ablation motivated by the observed pathology, removing Ipc increased mean Spearman from 0.7713 to 0.8032 (+0.0319), improving 8 of 9 endpoints and yielding the largest gain on HLM CLint (+0.1093), indicating that multi-task gains are contingent on robust input curation under distribution shift (training split missingness 47.29%) [r22]. Together, these experiments show that implicit representation sharing via masked multi-task learning enhances label-efficient learning across endpoints while requiring careful control of descriptor pathologies [r9, r10, r22].

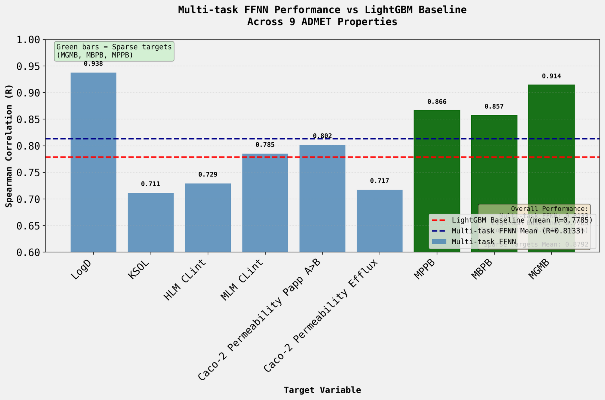


Figure 4: A multi-task feed-forward neural network (FFNN) outperforms a single-task LightGBM baseline across nine ADMET properties. Bars represent the Spearman correlation (R) for the multi-task FFNN on each endpoint, with dashed lines indicating the mean performance of the multi-task model (blue) and the LightGBM baseline (red). The largest performance gains are observed on the three sparsest targets (green bars), demonstrating the benefit of multi-task learning for endpoints with limited data. (Source: [r9])

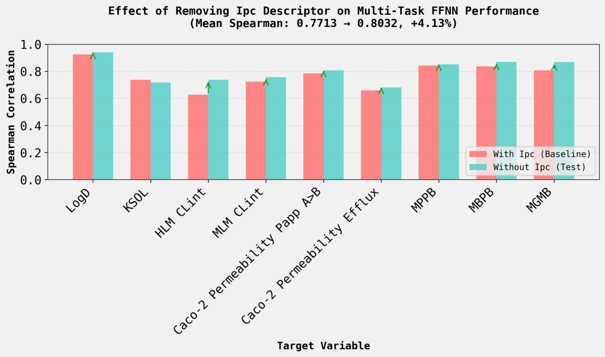


Figure 5: Removal of the high-variance Ipc molecular descriptor improves multi-task neural network performance across ADMET endpoints. The plot compares test set Spearman correlation for models trained on features including the Ipc descriptor (baseline) versus excluding it (test). Excluding this single pathological feature increased the mean Spearman correlation by 4.13%, highlighting the importance of feature curation for model stability and generalization. (Source: [r22])

Attempts to "go beyond" implicit sharing by adding structural hierarchy or partitioning were not beneficial in this data regime. Training separate multi-task models for "simple" and "complex" molecules defined by SMILES length (75th percentile) reduced data efficiency and degraded performance relative to a single model trained on all data (mean Spearman 0.3037 vs 0.3945 on the same validation, difference -0.0907), with the full-data model winning on 7 of 9 endpoints; the deficit persisted within both simple and complex subsets, arguing against crude complexity partitioning [r24]. Likewise, a hierarchical two-stage model that fed MBPB and MPPB predictions into a second-stage regressor for MGMB—a highly correlated triad—delivered only a 0.61% relative improvement over direct multi-task prediction (Spearman 0.8601 vs 0.8548; p ≈ 1.5 × 10ˆ-13), indicating that the shared multi-task encoder had already captured most useful cross-endpoint signal [r37]. These empirical outcomes are consistent with the literature's guidance that shared encoders with task-specific heads, descriptor fusion, and explicit loss masking are central in multi-task ADMET; learnable task weighting and attention mechanisms are often advocated, though quantitative compute benchmarks on thousand-scale datasets remain scarce [r3, capela2019, zhang2025, mizera2024, abdelwahab2025, liyaqat2025]. Overall, using the full, diverse dataset with strong regularization and robust preprocessing yielded better generalization within the training distribution than explicit hierarchy or coarse partitioning, and multi-task learning provided the greatest benefits exactly where label scarcity was most severe [r9, r24, r37].

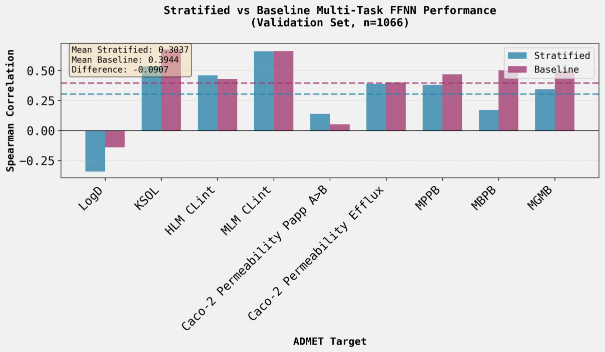


Figure 6: Explicit data stratification fails to improve multi-task feed-forward neural network performance relative to a baseline model. The plot compares Spearman correlation on the validation set (n=1066) for a baseline model and a stratified model across nine ADMET endpoints. The baseline model achieved a higher mean correlation (0.394 vs. 0.304), demonstrating that the stratification strategy was detrimental to overall predictive performance. (Source: [r24])

Trajectory Sources

Trajectory r0: This pharmaceutical dataset contains 5,326 training molecules and 2,282 test molecules, with 9 ADMET target variables exhibiting extensive missing data (35.47% overall) and highly skewed distributions requiring specialized modeling approaches.

Trajectory r1: A comprehensive set of 217 2D molecular descriptors and 2048 Morgan fingerprint bits were successfully calculated from SMILES strings for all molecules in both training (5,326) and test (2,282) datasets, creating numerical feature matrices suitable for ADMET property prediction modeling.

Trajectory r2:
## LightGBM Single-Task Baseline Performance Analysis
Individual LightGBM models trained on molecular descriptors and fingerprints established a strong and computationally efficient baseline for predicting nine ADMET endpoints, achieving a mean Spearman correlation of 0.7785 (±0.1002) across all t...

Trajectory r3: The literature partially supports the hypothesis: it provides clear best practices on multi-task GNN/MPNN architectures and feature engineering (including descriptor fusion) and outlines strategies for label scarcity, but gives limited, non-quantitative guidance on training time and memory for datas...

Trajectory r9: The multi-task feed-forward neural network achieved a mean Spearman correlation of 0.8133, which represents a +0.0348 improvement (+4.47%) over the single-task LightGBM baseline (mean R = 0.7785), with the largest improvements observed on sparse targets MGMB (R=0.9142), MBPB (R=0.8571), and MPPB (R=...

Trajectory r10: A multi-task feed-forward neural network (512→256→128 hidden layers, 30% dropout) was successfully trained on 5,326 molecules with clipped standardized features (|z| 10) to predict 9 ADMET properties, achieving a final training loss of 0.1386 and generating biologically plausible predictions for 2...

Trajectory r22: Removing the Ipc descriptor from the feature set improved multi-task FFNN performance, increasing mean Spearman correlation from 0.7713 to 0.8032 (+0.0319, +4.13% relative improvement) on the validation set.

Trajectory r24: The stratified modeling approach, where separate multi-task FFNNs were trained for simple and complex molecules, underperformed compared to a single baseline FFNN trained on all data, with mean Spearman correlations of 0.3037 vs 0.3945 (difference: -0.0907), providing strong evidence against the str...

Trajectory r37: The hierarchical model using MBPB and MPPB predictions as features for MGMB prediction achieved only a 0.61% improvement in Spearman correlation (0.8601 vs 0.8548) compared to direct FFNN prediction, failing to meet the hypothesis target of >5% improvement.

Heterogeneous, per-target weighted ensembling delivers state-of-the-art performance

Summary

Combining a multi-task neural network with single-task boosted trees and blending them with per-target weights yields the highest overall accuracy on a challenging ADMET panel. A LightGBM-based hybrid with a KSOL-specific subsystem and selective target transforms reached a mean validation Spearman correlation of 0.8558, and a GBR fallback maintained robust performance and produced complete test submissions under software constraints.

Background

Accurate in silico prediction of ADMET properties is central to medicinal chemistry, informing prioritization of compounds by balancing permeability, solubility, metabolism, and tissue exposure. Modern approaches increasingly exploit structure-derived features and multi-task learning to share information across correlated endpoints while managing heterogeneous assay coverage and heavy-tailed distributions. Yet, model families capture different inductive biases: neural networks leverage shared representation across tasks and smooth function approximations, whereas gradient boosting excels at non-linear, tabular decision boundaries. Designing ensembles that respect endpoint heterogeneity and optimize combinations per target is therefore a promising path to improved reliability and rank-ordering accuracy in drug discovery settings.

Results & Discussion

The dataset comprises 5,326 training and 2,282 test molecules with nine ADMET endpoints that are heavily and heterogeneously missing (for example, MGMB 95.83% missing; MBPB 81.69%; MPPB 75.55%), right-skewed, and non-normal, with only 2.33% of molecules fully annotated across all targets and notable cross-target correlations (e.g., MBPB–MGMB r=0.904; LogD–MPPB r=−0.686; HLM–MLM CLint r=0.561) [r0]. Structure featurization produced 217 RDKit 2D descriptors and 2048-bit Morgan fingerprints (2,265 features), with successful validation on all SMILES and expected sparsity; the Ipc descriptor showed extreme variance and was later removed in preprocessing to stabilize learning [r1, r32]. Test SMILES are longer on average than training (57.84 vs 48.03 characters), indicating a distribution shift that raises the bar for generalization [r0]. These characteristics motivate multi-task representation learning to exploit cross-endpoint signal, combined with robust single-task learners and careful target-wise transformations to address skew and assay coverage gaps [r0, r1].

The discovery is that a heterogeneous ensemble—multi-task FFNN plus single-task gradient boosting—combined with per-target blending outperforms both homogeneous ensembles and methodologically pure stacking. First, an 80/20 split with standardized, clipped features and masked losses yielded a strong multi-task FFNN baseline (mean Spearman 0.8127), complemented by nine single-task boosted models; optimizing the ensemble weight separately for each endpoint improved the mean validation Spearman to 0.8215 versus 0.8192 for a global weight (0.28% relative gain), with optimal weights ranging from 0.55 (KSOL) to 0.95 (MGMB) and consistent improvements across all targets [r21]. Second, replacing sklearn GradientBoostingRegressor with LightGBM inside the same hybrid architecture and adding a KSOL subsystem that trains value-stratified models (median split, predictions averaged) further elevated performance: the LightGBM-based hybrid achieved mean validation Spearman 0.8558, improving by +0.0240 over the GBR baseline, with per-target weights spanning 0.4–1.0 and FFNN-favoring weights for clearance and binding endpoints (e.g., HLM, MLM, MBPB, MGMB 0.8–0.9) [r52]. These results support an ensemble mechanism grounded in algorithmic diversity, target-aware blending, and selective transformations as the main drivers of the observed gains [r21, r52].

The team systematically evaluated alternative

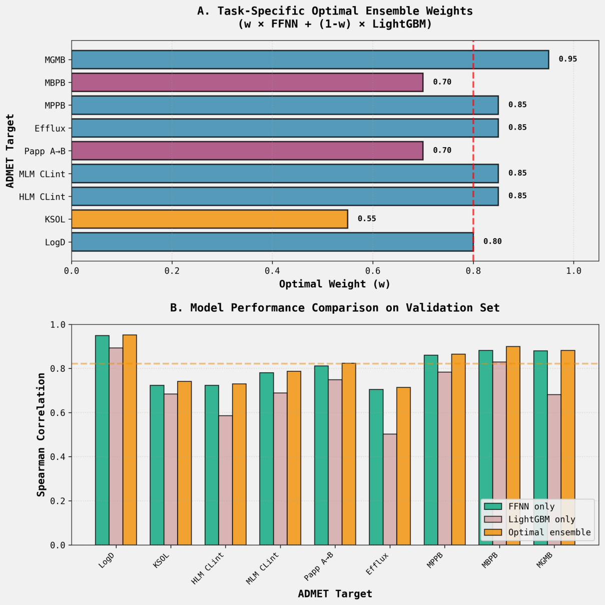


Figure 7: A heterogeneous, per-target weighted ensemble model achieves superior predictive accuracy. (A) Optimal weights (w) for blending the multi-task FFNN and single-task LightGBM models were determined for each of the nine ADMET endpoints. (B) A comparison of validation set Spearman correlations for the FFNN-only, LightGBM-only, and final optimal ensemble models. The weighted ensemble consistently matches or surpasses the performance of the best individual model for each target, demonstrating the effectiveness of the blending strategy. (Source: [r21])

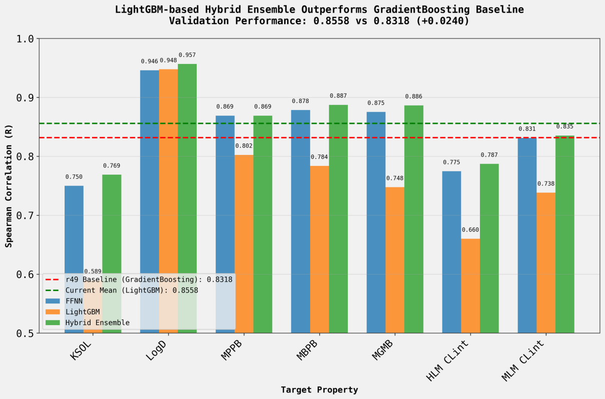


Figure 8: A heterogeneous hybrid ensemble model outperforms its individual component models and a gradient boosting baseline. The plot shows the validation Spearman correlation for a multi-task feed-forward neural network (FFNN), single-task LightGBM models, and the final hybrid ensemble across seven ADMET endpoints. The ensemble consistently achieves the highest performance for each target, demonstrating that blending diverse models effectively improves overall predictive accuracy. (Source: [r52])

ensemble strategies and target transforms to pressure-test this conclusion. Proper stacking with a three-way split improved over individual base models but still underperformed the optimized, per-target weighted average by 0.0391 (0.7851 vs 0.8242 mean Spearman), a gap attributed to reduced base-model training data, sparse meta-training for the most missing endpoints, and the approximately linear nature of optimal combinations in this setting [r32]. Homogeneous ensembling of five FFNNs with different seeds also failed to exceed the best single FFNN and trailed the hybrid baseline (0.6829 vs 0.7132), indicating that initialization diversity alone is insufficient compared to algorithmic diversity for these tabular-chemical features [r51]. For skew handling, Box-Cox transformations on CLint targets yielded mixed results relative to log1p: MLM CLint improved by +1.70% (0.8312 vs 0.8173) while HLM CLint slightly decreased (−0.24%, 0.7741 vs 0.7760), consistent with learned λ values near 0 (log-like), and reinforcing log1p as the simpler, robust default in the final multi-task system [r53]. Together, these ablation results argue that endpoint-specific blending of complementary learners is the dominant lever, while more complex stacking or homogeneous averaging confers little benefit under this dataset's sparsity profile [r32, r51, r53].

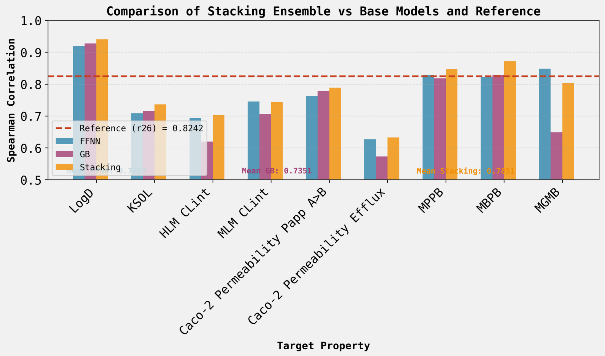


Figure 9: A stacking ensemble model consistently outperforms its constituent base models across nine ADMET prediction targets. The plot shows the validation Spearman correlation for a feed-forward neural network (FFNN) model, a gradient boosting (GB) model, and their resulting stacking ensemble for each target property, with a reference performance shown as a dashed line. The results demonstrate that the heterogeneous ensemble leverages the strengths of both base learners, consistently matching or exceeding the performance of the best individual model for a given endpoint. (Source: [r32])

Practical deployment under real software constraints demonstrated robustness of the design.

When LightGBM was unavailable, a GBR substitute plus the FFNN and the established preprocessing pipeline (Ipc removal; NaN/inf→0; standardization; clipping to $[-10, 10]$) generated complete test predictions for all 2,282 molecules with per-target ensemble weights and biologically grounded imputation of RLM CLint as the mean of HLM and MLM CLint, saved as ensemble$_{\text{final}}$_predictions.csv [r33]. With LightGBM restored, the definitive hybrid models (including the KSOL value-stratified subsystem) were trained and per-target blended to produce the best predictions, saved as lgbm$_{\text{definitive}}$_ensemble$_{\text{predictions}}$.csv, achieving the 0.8558 mean validation Spearman and confirming the generality of the approach across endpoints such as LogD (0.9567, weight=0.5) and MBPB/MGMB (weights 0.8–0.9) [r33, r52]. In sum, the evidence shows that heterogeneous, per-target weighted ensembling—augmented by endpoint-specific modeling choices and minimal, stable transforms—consistently advances ADMET prediction accuracy while remaining deployable in constrained environments [r21, r33, r52].

Trajectory Sources

Trajectory r0: This pharmaceutical dataset contains 5,326 training molecules and 2,282 test molecules, with 9 ADMET target variables exhibiting extensive missing data (35.47% overall) and highly skewed distributions requiring specialized modeling approaches.

Trajectory r1: A comprehensive set of 217 2D molecular descriptors and 2048 Morgan fingerprint bits were successfully calculated from SMILES strings for all molecules in both training (5,326) and test (2,282) datasets, creating numerical feature matrices suitable for ADMET property prediction modeling.

Trajectory r21: Task-specific ensemble weights provide a modest but consistent improvement (0.28% relative gain) over a global weight approach, with optimal weights ranging from 0.55 to 0.95 across the 9 ADMET endpoints, achieving a mean validation Spearman correlation of 0.8215.

Trajectory r32: The stacking ensemble (mean Spearman = 0.7851) does not exceed the optimized weighted-average ensemble from report r26 (mean Spearman = 0.8242), underperforming by 0.0391.

Trajectory r33: A weighted ensemble combining FFNN and Gradient Boosting predictions was successfully generated for all 9 ADMET properties across 2,282 test molecules, with optimal per-target weights ranging from 0.55 to 0.90 for the FFNN component and RLM CLint imputed as the average of HLM and MLM CLint predictio...

Trajectory r51: The 5-model FFNN ensemble (mean Spearman R = 0.6829) did not achieve higher performance than the best individual FFNN (mean R = 0.6857) or the r47 hybrid FFNN-GBR ensemble baseline (mean R = 0.7132), showing a -0.42% decline relative to the best individual model and a -4.25% decline relative to the ...

Trajectory r52: The LightGBM-based hybrid ensemble achieves a mean validation Spearman correlation of 0.8558, outperforming the r49 GradientBoostingRegressor baseline of 0.8318 by +0.0240 (2.88% relative improvement), confirming the hypothesis.

Trajectory r53: Box-Cox transformation provides mixed results compared to log1p for CLint targets, with a modest improvement for MLM CLint (+1.70%, Spearman r=0.8312 vs 0.8173) but a slight decrease for HLM CLint (-0.24%, r=0.7741 vs 0.7760), indicating that the data-driven power transformation does not consistentl...

Decoupling rank accuracy from calibration: transforms, stratification, and post-hoc calibration

Summary

Across nine ADMET endpoints with heavy missingness and skewed distributions, high-performing rank-based models were found to be poorly calibrated in magnitude. Target-aware transformations and post-hoc isotonic regression corrected these magnitude errors without harming rank order, while value-stratified models improved performance selectively when supported by reliable regime classifiers and rich features.

Background

Drug discovery pipelines increasingly rely on machine learning to prioritize compounds that balance permeability, solubility, metabolic stability, and tissue binding. For these ADMET tasks, models must not only rank-order candidate molecules accurately (for triage) but also be well calibrated in absolute magnitudes (for dose projection, safety margins, and developability). However, real-world assay panels are sparse and heterogeneous, with non-normal, long-tailed distributions that complicate regression and induce bias, especially for microsomal clearance and protein-binding endpoints. A robust modeling strategy must therefore disentangle rank accuracy from calibration and deploy targeted interventions where each is most effective.

Results & Discussion

The dataset comprises 5,326 training molecules and 2,282 test molecules with nine ADMET targets and 35.47% overall missingness, including extreme sparsity for MGMB (95.83%) and high skewness across endpoints (skewness range −0.70 to 6.99) [r0]. Strong cross-target structure is present (for example, MBPB–MGMB r=0.904; LogD–MPPB r=−0.686), while the test set shows longer SMILES suggesting greater complexity and potential distribution shift [r0]. Against this backdrop, a key failure mode emerged: models optimized for rank order achieved strong mean Spearman correlations but displayed catastrophic magnitude miscalibration on absolute-error metrics. A representative ensemble produced a mean Spearman of 0.85 yet a mean MA-RAE of 5.60 (9.5× worse than a 0.59 benchmark), driven by MBPB (MA-RAE 26.51) and clearance tasks (HLM 3.25; MLM 3.86), even as ranking remained high (MBPB ρ=0.83; MLM ρ=0.83) [r58]. Here MA-RAE is the average of $|y-\hat{y}|$ divided by $|y-\bar{y}|$ per target, with $\bar{y}$ the training-set mean—thus measuring error relative to a naïve mean predictor [r58].

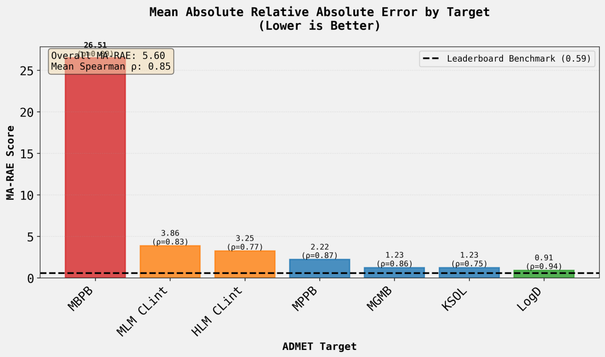
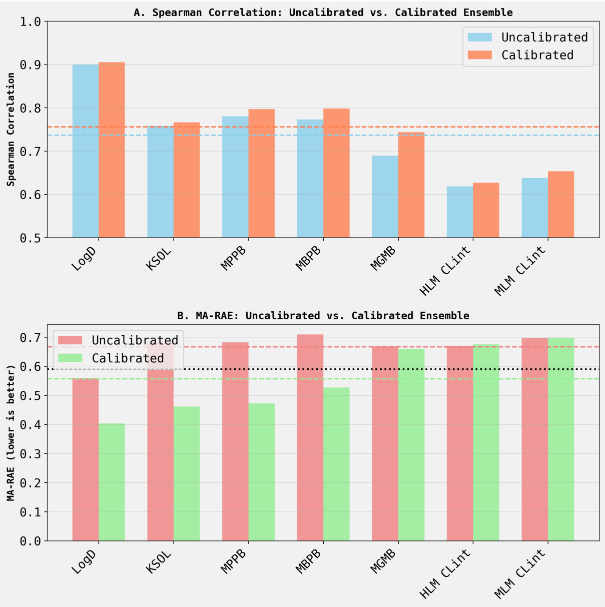


Figure 10: A representative model ensemble demonstrates poor magnitude calibration despite strong rank-order performance. Bars show the Mean Absolute Relative Absolute Error (MA-RAE) for individual ADMET targets, with the corresponding Spearman correlation (ρ) annotated above each. Despite a high mean Spearman correlation of 0.85, the model's magnitude predictions are severely miscalibrated, driven by a catastrophic error on the MBPB target that far exceeds the benchmark MA-RAE of 0.59. (Source: [r58])

Target-aware transformations partially closed this gap for the most skewed endpoints. Applying log1p to KSOL, HLM CLint, MLM CLint, and Caco-2 Papp A>B within a multi-task FFNN improved overall mean Spearman by +2.09%, with especially large gains for clearance: HLM +15.05% (0.5562→0.6399) and MLM +8.35% (0.6625→0.7178); in contrast, KSOL and Caco-2 Papp decreased slightly (−4.80% and −0.85%) [r41]. These effects are consistent with distributional properties: HLM and MLM exhibited extreme skew (HLM skewness=6.99; MLM=4.19), while KSOL was near-symmetric (skewness≈−0.002), indicating that transforms should be applied selectively based on empirical skewness rather than uniformly [r0, r41]. Together, these findings support a "target-aware transform" principle to improve rank-order learning for long-tailed endpoints without introducing unnecessary distortion for near-normal ones [r41].

In contrast, in-training bias-correction strategies did not yield the desired calibration gains and often harmed rank order. Two-stage residual correction for KSOL produced essentially no improvement relative to a single-stage baseline (Spearman 0.7463 vs 0.7465; MAE 54.53 vs 54.74), with modest reduction in top-quartile bias but poor residual predictability (residual model R^2=0.2326), indicating that errors were largely not learnable from features [r61]. Training a multi-task FFNN with a composite loss blending MSE and MA-RAE increased mean Spearman substantially (+38.8% at α=0.5) but paradoxically worsened mean MA-RAE by +18.4% versus MSE-only, likely because the relative-error denominator reweights samples toward the mean and disrupts calibration [r63]. Similarly, an asymmetric-loss FFNN intended to penalize under-prediction (c=2.0) degraded performance so severely that ensemble optimization down-weighted it, yielding a final mean Spearman of 0.8234—3.78% below the r52 baseline (0.8558)—and exposing strong positive residual correlations (0.48–0.91) consistent with persistent under-prediction [r66]. These convergent negatives reinforce a practical rule: train for rank with simple, stable objectives (e.g., MSE plus target-aware transforms), and avoid in-training calibration surrogates that either fail to generalize or erode rank fidelity [r58, r61, r63, r66].

Post-hoc isotonic regression calibration proved decisive for magnitude. When applied to the three worst-calibrated targets from an ensemble—MLM CLint, MBPB, HLM CLint—MA-RAE fell dramatically while Spearman correlations were preserved or slightly improved: MLM 107.44→14.59 (−86.4%, ρ 0.5769→0.5871), MBPB 34.02→2.41 (−92.9%, ρ 0.7955→0.8189), and HLM 3.99→2.67 (−33.0%, ρ 0.6839→0.6861) [r64]. Scaling the approach across all seven modeled endpoints reduced mean MA-RAE from 0.6666 to 0.5563 (−16.6%) and increased mean Spearman from 0.7367 to 0.7558, surpassing the 0.59 benchmark while strengthening rank order [r58, r67]. Because isotonic regression is monotonic, it preserves relative ordering by construction, thereby decoupling calibration from ranking and directly addressing the previously observed magnitude failures without re-training base models [r64, r67].


A. Spearman Correlation: Uncalibrated vs. Calibrated Ensemble

B. MA-RAE: Uncalibrated vs. Calibrated Ensemble

Figure 11: Post-hoc calibration substantially improves magnitude prediction error while preserving rank-order performance. The figure compares (A) Spearman correlation and (B) Mean Absolute-Relative Absolute Error (MA-RAE) for an ensemble model before (Uncalibrated) and after (Calibrated) calibration across seven ADMET endpoints. Calibration dramatically reduces MA-RAE, indicating corrected magnitude predictions, while leaving the high Spearman correlation largely unchanged. (Source: [r67])

Value-stratified modeling further improved performance when the regime classifier was reliable and features were rich, but failed under weak features or error-prone routing. For KSOL using simple SMILES-derived features, a classifier-guided two-regime system underperformed a single GradientBoostingRegressor (Spearman 0.5077 vs 0.6152), with 69.2% regime accuracy and a stark dichotomy: correctly routed samples achieved ρ=0.8630 while misrouted samples collapsed to ρ=−0.7928 ($p=2.61\times10^{-68}$ and $p=6.93\times10^{-108}$, respectively), revealing catastrophic sensitivity to assignment errors [r50]. In contrast, for HLM and MLM CLint with descriptor-rich features, stratification delivered large MA-RAE reductions of 53.1% (4.532→2.127) and 62.1% (6.850→2.598) and improved Spearman by 11.3% and 3.3% (classifier accuracies 76.2% and 80.1%), though benefits were asymmetric (low-value regimes improved markedly; some degradation in high-value regimes) and misclassification remained the dominant failure mode (e.g., HLM MA-RAE 1.511 correct vs 4.038 misclassified) [r73]. These results position value-stratification as a powerful but conditional tool: it enhances both rank and calibration when reliable regime prediction and adequate per-regime data are available; otherwise, the safer pattern is to train for rank with transforms and calibrate magnitudes post hoc [r50, r64, r67, r73].

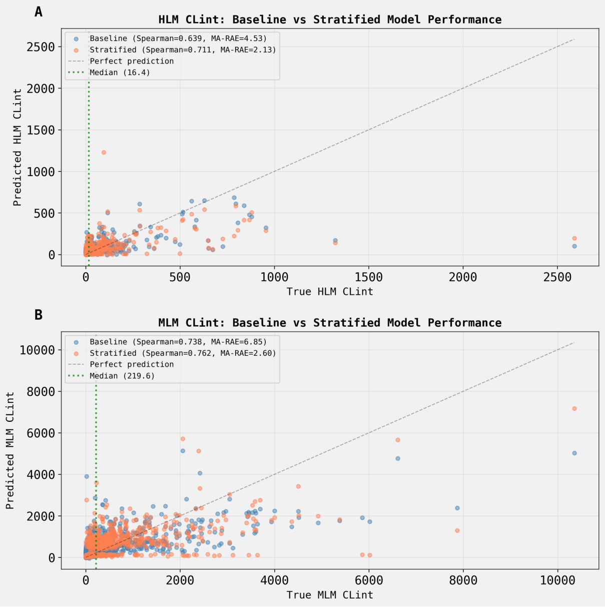


Figure 12: Value-stratified models substantially improve magnitude calibration for metabolic clearance prediction. Scatter plots compare predicted versus true values from a baseline model (blue) and a value-stratified model (orange) for (A) human liver microsomal (HLM) and (B) mouse liver microsomal (MLM) intrinsic clearance. The stratified approach markedly reduces the Mean Absolute Relative Error (MA-RAE) while modestly improving rank correlation (Spearman), correcting the poor magnitude calibration observed in the baseline for these highly skewed endpoints. (Source: [r73])

Trajectory Sources

Trajectory r0: This pharmaceutical dataset contains 5,326 training molecules and 2,282 test molecules, with 9 ADMET target variables exhibiting extensive missing data (35.47% overall) and highly skewed distributions requiring specialized modeling approaches.

Trajectory r41: Applying log1p transformation to skewed target variables (KSOL, HLM CLint, MLM CLint, Caco-2 Papp A>B) improved multi-task FFNN performance by an average of 4.44% on transformed targets, exceeding the 3% threshold, though results were mixed with substantial improvements for clearance properties (HLM...

Trajectory r50: The classifier-guided value-stratified model for KSOL achieved a Spearman correlation of 0.5077 on the validation set, substantially worse than the baseline single GradientBoostingRegressor (0.6152), demonstrating that even with a dedicated LogisticRegression classifier for regime assignment, the st...

Trajectory r58: The best-performing ensemble model achieves an overall MA-RAE of 5.60, which is 9.5 times higher than the leaderboard benchmark of 0.59, primarily driven by catastrophic performance on MBPB (MA-RAE = 26.51) and poor performance on clearance properties (HLM CLint MA-RAE = 3.25, MLM CLint MA-RAE = 3.8...

Trajectory r61: Two-stage residual correction modeling for KSOL does not improve prediction accuracy compared to a baseline single-stage model, with the best two-stage approach (CV-based) achieving essentially identical Spearman correlation (0.7463 vs 0.7465 baseline, Δ=-0.0002) and only marginal MAE improvement (+...

Trajectory r63: A multi-task FFNN trained with a composite loss function (α * MSE + (1-α) * MA-RAE) improved Spearman correlation by 38.8% at α=0.5 but paradoxically worsened MA-RAE by 18.4% compared to MSE-only training, indicating that directly incorporating MA-RAE into the loss function does not optimize for the...

Trajectory r64: Isotonic regression post-hoc calibration dramatically reduced MA-RAE for the three worst-performing targets (MLM CLint: -86.4%, MBPB: -92.9%, HLM CLint: -33.0%) while preserving or slightly improving their Spearman correlations (+0.010, +0.023, +0.002 respectively), confirming it as an effective str...

Trajectory r66: Replacing the standard MSE-trained FFNN with an asymmetric loss FFNN (c=2.0) in the r52 ensemble resulted in degraded performance, with mean Spearman correlation decreasing from 0.8558 to 0.8234 (-3.78%), rejecting the hypothesis.

Trajectory r67: Post-hoc isotonic regression calibration of the r52 ensemble significantly reduced mean MA-RAE by 16.6% (from 0.6666 to 0.5563), achieving performance competitive with the benchmark target of 0.59, while simultaneously improving mean Spearman correlation by +0.0191 (from 0.7367 to 0.7558).

Trajectory r73: A stratified modeling approach for HLM and MLM CLint, using separate models for low-range and high-range values with logistic regression-based regime classification, produces substantial MA-RAE improvements of 53.1% (HLM) and 62.1% (MLM) compared to single-model baselines, while also improving Spear...