Cross-species multi-task learning with molecular and ADME descriptors for liver microsomal metabolic stability

Subhin Seomun[1] and Sunyong Yoo[1, 2*]
[1]Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju, Republic of Korea

[2]R&D Center, MATILO AI Inc., Gwangju, Republic of Korea

[*]Correspondence: syyoo@jnu.ac.kr; Tel.: +82-62-530-1810

Abstract

Liver microsomal stability is a key determinant of *in vivo* compound exposure and efficacy. Although metabolic stability has been extensively studied, linking substructure destabilizing features to absorption, distribution, metabolism, and excretion (ADME) properties remains challenging. Moreover, single-species, single-modality models often generalize poorly. To address these limitations, we propose a cross-species multi-task learning framework that integrates multi-modal molecular representations to predict liver microsomal stability. Specifically, the model leverages three complementary modalities: SMILES-derived fingerprints, molecular graphs, and *in silico* ADME descriptors. These modalities are learned in a shared network using data from multiple species and subsequently fused via attention mechanisms to form a shared molecular representation, which captures conserved structure-metabolism relationships common across species. Species-specific network capture individual metabolic characteristics and stability predictions for human (HLM), rat (RLM), and mouse liver microsomal (MLM). Under stratified 10-fold cross-validation, mean AUROC was 0.770 ± 0.001 (HLM), 0.785 ± 0.001 (RLM), and 0.766 ± 0.001 (MLM). To understand the chemical basis of metabolic liability, we examined three multi-level perspectives. At the molecular property level, physicochemical descriptors related to enzyme interaction, permeability/transport, and the lipophilicity-polarity axis emerged as dominant predictive drivers. At the substructure level, to pinpoint specific sites of metabolic vulnerability, recurring destabilizing features were identified at alkenes and allylic/benzylic positions, while amide and carbamate carbonyl motifs conferred stability. To elucidate the underlying physicochemical mechanisms, these structural motifs were linked to systematic shifts in logP, solubility, blood-brain barrier propensity, and efflux liability. Overall, these results indicate that the cross-species integrative model accurately predicts microsomal stability across human, rat, and mouse while providing chemically grounded explanations.

Keywords: Liver microsomal metabolic stability; Multi-modal; ADME; Multi-task learning;

1. Introduction

Liver microsomal stability is central to drug discovery, guiding lead optimization by predicting biotransformation and transport relevant to *in vivo* exposure [1-4]. In practice, human (HLM), rat (RLM), and mouse (MLM) microsomal assays are frequently used in tandem to guide early structure optimization and to anticipate species differences that can complicate translation from discovery to preclinical studies [1, 5]. Connecting chemical destabilizing features at the fragment level specific substructures prone to metabolic instability or toxicity to overall ADME descriptors remains challenging [6, 7]. Establishing this connection is critical for rational drug design, as it enables medicinal chemists to identify and modify problematic substructures early in development while simultaneously predicting their impact on whole-molecule pharmacokinetic properties across species. Moreover, metabolic stability reflects complex biotransformation pathways involving multiple enzymes and metabolites. This complexity limits the utility of simple empirical rules, motivating data-driven approaches[2, 8, 9].

Computational approaches have addressed this challenge through multiple strategies. MetStabOn developed species-specific models for human, rat, and mouse liver microsomal using ligand-based features to classify compounds into low, medium, and high stability categories [10]. PredMS built a Random Forest binary classifier for HLM (≥50% remaining at 30 min) and reported moderate performance on an external set [11]. CMMS-GCL integrates SMILES-based and graph-based encoders with contrastive learning to improve representation quality under data scarcity [12]. To address cross-species prediction, Long et al. [13] unrated a large-scale HLM/RLM/MLM dataset and developed descriptor-based and graph neural network models. Using SHAP and atom-level attribution, they identified both shared and species-specific metabolic determinants, demonstrating the value of interpretable cross-species modeling. The utilization of computational modeling has facilitated the prioritization of stable chemotypes. However, many existing models rely on single-modality structural encodings and single-species datasets. As a result, such models may underutilize complementary sources of information and often provide limited mechanistic insight into why a molecule is predicted to be stable or unstable [11, 12, 14]. Moreover, the majority of frameworks do not explicitly integrate *in silico* ADME descriptors, such as permeability surrogates, drug-likeness rule indices, and enzyme-interaction flags, alongside structural encodings. This integration is essential for translating fragment-level modifications into predictions of whole-molecule pharmacokinetic behavior, yet remains underexplored in current approaches [12, 14]. Finally, model interpretability is typically summarized as feature rankings at the molecular level direct attribution to chemically meaningful substructures and the alignment of those substructures with ADME descriptor across species have been less explored [15].

To address these gaps, we propose a cross-species multi-task learning framework that integrates multi-modal molecular representations to predict liver microsomal stability. Specifically, the model leverages three complementary modalities: SMILES-derived fingerprints, molecular graphs, and *in silico* ADME descriptors. Multi-task learning enables the model to exploit correlations among species while preserving species-specific metabolic differences [16]. These representations are processed within a shared network and species-specific independent networks, a design that captures generalizable structure-metabolism relationships while allowing species-dependent effects to be expressed through dedicated output layers. The model is trained and evaluated on a curated cross-species liver microsomal stability dataset from PubChem BioAssay [17]. This study has two main objectives. First, we seek to enhance cross-species prediction of microsomal stability through a multi-task learning framework that integrates structural and ADME features using multi-modal molecular representations. Second, we identify the substructures driving each prediction and link them to

ADME descriptors commonly used in medicinal chemistry.

2. Materials and Methods

2.1 Dataset collection

The liver microsomal stability data were obtained from the PubChem BioAssay database [18]. We collected 6,703 HLM, 5,763 RLM, and 6,501 MLM measurements, for a total of 18,967 records. We excluded entries lacking half-life measurements as well as data from assays conducted in the presence of specific CYP isoform inhibitors [19, 20]. Compounds with half-life $t_{1/2} > 30$ min were labeled stable, while those with $t_{1/2} < 30$ min were labeled unstable [1, 21]. Compounds with half-life values of exactly 30 min were excluded to avoid ambiguity at the classification boundary. We curated the data by removing entries with missing or invalid SMILES strings, standardizing molecular structures using RDKit through neutralization and salt removal, and filtering out inorganic compounds [22]. The dataset was split into training (80%) and test (20%) sets. Figure 1 shows stability classes in the training and test sets for each species and the degree of compound overlap across species.

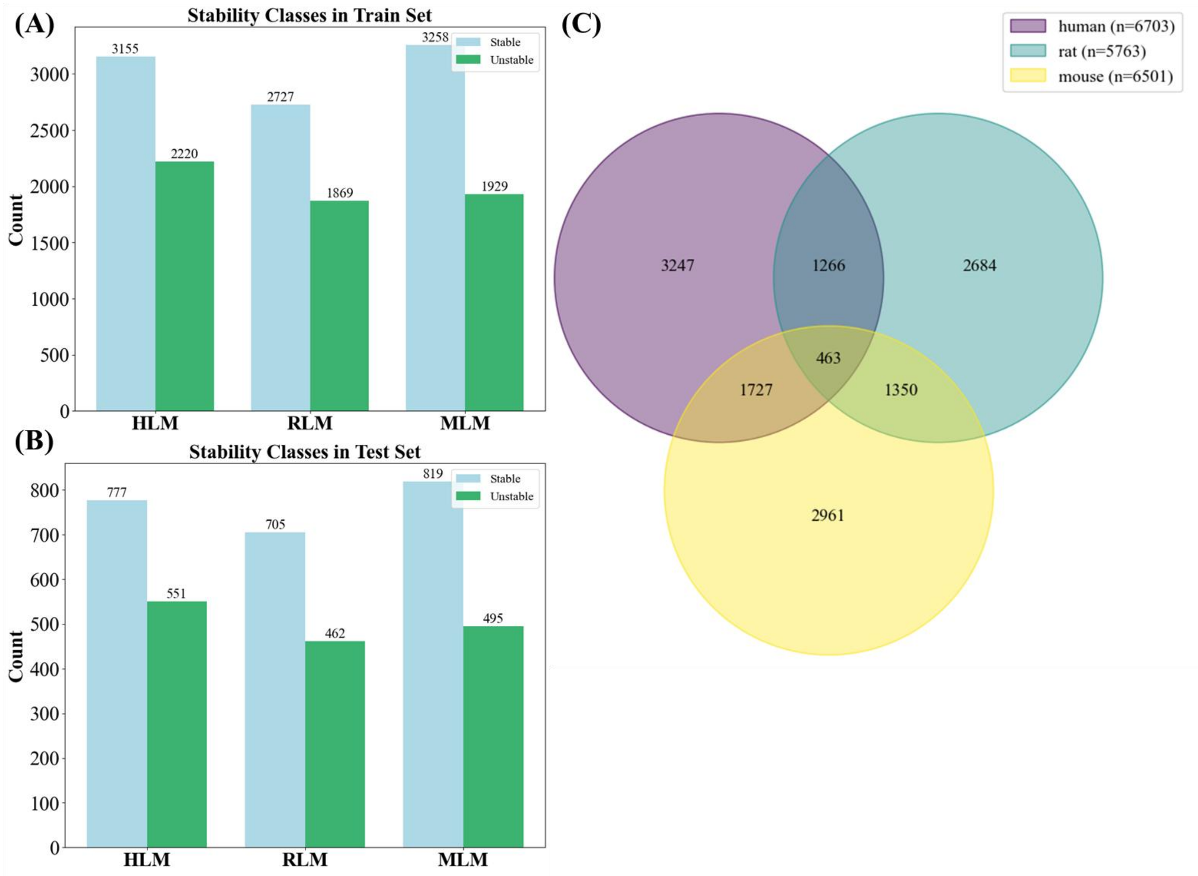

Fig. 1. Composition of the microsomal stability dataset and cross-species overlap. The number of compounds labeled as stable and unstable in (A) training and (B) test sets for HLM, RLM, and MLM. (C) Venn diagram of unique SMILES across species after filtering and structure standardization.

2.2 Method overview

The model predicts metabolic stability for HLM, RLM, and MLM through a multi-modal, multi-task architecture consisting of a shared network and species-specific independent networks (Fig 2). Three molecular representations serve as input (Fig. 2A): SMILES-derived fingerprints including Morgan, MACCS, and RDKit, molecular graphs with node and edge features, and in silico ADME descriptors. In the shared network, modality-specific embeddings are generated through fingerprint concatenation, a GCN block for graph processing, and MLP blocks for descriptor encoding (Fig. 2B). These embeddings are then concatenated through a feature fusion layer in the independent networks (Fig. 2C), where the fused representation is passed to species-specific networks to model individual metabolic characteristics. Finally, each species-specific network generates binary stability predictions (Fig. 2D). This architecture captures generalizable structural features through the shared network while modeling species-specific metabolic characteristics through dedicated independent networks.

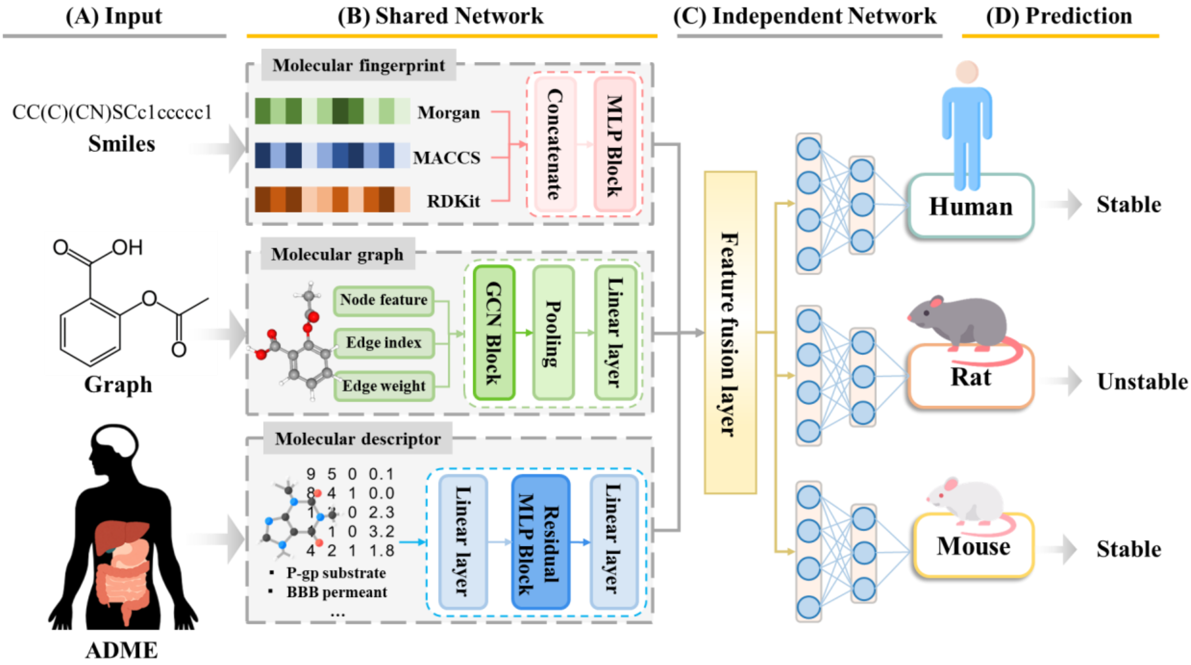

Fig. 2. Method overview of the proposed multi-modal, multi-task framework for liver microsomes stability prediction. (A) Three molecular representations are used as input, derived from SMILES, molecular graph, and ADME descriptors. (B) Shared network that encodes three molecular representations. Morgan, MACCS, and RDKit fingerprints derived from SMILES are processed through an MLP block. Molecular graphs with node and edge features pass through a GCN block with pooling. *In silico* ADME and physicochemical descriptors are processed through stacked linear layers and a residual MLP block. The resulting embeddings are combined in a feature fusion layer. (C) Species-specific independent networks receive the fused representation for human rat and mouse. (D) Each prediction outputs a binary prediction of Stable or Unstable for the corresponding species.

2.3 Molecular representation and embedding

2.3.1 Molecular fingerprint

The molecular structure is represented by three complementary two-dimensional fingerprints: Morgan [23], MACCS keys [24], and the RDKit [25] topological fingerprint. Each fingerprint captures distinct aspects of molecular structure and connectivity. Morgan fingerprints encode circular atom neighborhoods by hashing local topologies, a representation widely used in structure-activity modeling. MACCS keys provide a 167-bit structural vector, where each bit indicates the presence or absence of a predefined substructure or functional group. This representation provides a compact, interpretable encoding that complements the hashed fingerprints[26]. RDKit topological fingerprints encode linear atom-bond paths of varying lengths, capturing connectivity patterns complementary to circular neighborhoods. For each molecule x, we compute the 2048-bit Morgan $M(x)$, 167-bit MACCS $A(x)$, and 2048-bit RDKit $R(x)$ fingerprints and concatenate them into a single vector:

$$Z(x) = M(x), A(x), R(x) \tag{1}$$

Let $Z(x)\in(0,\ 1)^{L}$ denote the concatenated fingerprint vector. If any fingerprint is excluded, L is reduced by the corresponding bit length. To obtain a task-relevant representation, we apply a two-layer (MLP) $f(Z(x);\ W)$ to the fingerprint vector. The MLP is defined as

$$f(Z(x); W) = \mathrm{ReLU}(W_2 * \mathrm{ReLU}(W_1 * Z(x) + b_1) + b_2) \tag{2}$$

This process reduces dimensionality while retaining informative structure. Where W_1 and W_2 are weight matrices, b_1 and b_2 are bias vectors with. $ReLU(z) = max\ (0, z)$. The MLP layers enable nonlinear feature transformations that capture complex structural patterns.

2.3.2 Molecular graph

The molecular topology is encoded by representing each molecule as an undirected graph $G = (V, E)$, where vertices represent atoms and edges represent covalent bonds [27]. Each node is represented by a 77-dimensional feature vector encoding: atom type in 55 categories, degree in 7 levels, total hydrogens in 6 levels, formal charge in 3 levels, hybridization in 5 types, and aromaticity as one binary feature. We employed a GCN-based graph module to learn molecular graph representations [28]. GCN aggregates neighborhood information through degree-normalized message passing. Let X be the node-feature matrix, $\tilde{A}=A+I$ the adjacency with self-loops, and $\widetilde{D}$ the diagonal degree matrix of $\tilde{A}$ be its degree matrix. The normalized propagation matrix is:

$$\hat{A} = \widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}} \tag{3}$$

With $H^{(0)} = X$, a GCN layer updates node embedding as:

$$H^{(l+1)} = \sigma\big(\hat{A}H^{(l)}W^{(l)}\big) \tag{4}$$

Where $W^{(l)}$ is a learnable weight matrix and σ is a nonlinear activation function. We stack two GCN layers, each followed by batch normalization and LeakyReLU activation. Dropout is applied after the first layer. A global max pooling then aggregates node embeddings into a graph-level vector, which is finally passed through a linear layer to yield a 256-dimensional graph embedding used by the downstream fusion module.

2.3.3 Molecular ADME & physicochemical descriptors

ADME descriptors were computed for each molecule using SwissADME [29]. We organized these descriptors into numeric and categorical features. Numeric features include physicochemical properties and drug-likeness metrics, including molecular weight, topological polar surface area (TPSA), lipophilicity estimates (iLOGP, XLOGP3, WLOGP, MLOGP), counts of heavy and aromatic atoms, the fraction of sp^3 hybridized carbons, the number of rotatable bonds, hydrogen-bond acceptor and donor counts, molar refractivity, and a skin permeation estimate. Additional features include violation counts for drug-likeness rules (Lipinski, etc.), bioavailability scores, PAINS and Brenk structural alerts, lead-likeness violations, and synthetic accessibility scores. Missing values were imputed with column means. Categorical features include gastrointestinal absorption, blood-brain barrier permeability, P-glycoprotein substrate status, and inhibition predictions for five cytochrome P450 isoforms (CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP3A4). Categorical features were one-hot encoded.

We encode the ADME descriptors x using a residual MLP with skip connections for stable training [30]. Initially, the input is subjected to a process of sanitization and normalization to yield $\hat{x}$=*norm*(*sanitize*(x)), then project it into a width stem $h^{(0)} = \sigma(W^{(s)}\ \hat{x}+b^{(s)})$, where $\sigma(\cdot)$ denotes a pointwise nonlinearity. The encoder consists of L residual blocks, each defined as:

$$l = 0, \dots, L - 1, \tag{5}$$

$$h^{(l+1)} = h^{(l+1)} + G_l\big(h^{(l)}\big), \tag{6}$$

$$G_l(h) = \ W_l^{(2)}\sigma\Big(W_l^{(1)} norm(h) + b_l^{(1)}\Big) + b_l^{(2)} \tag{7}$$

This formulation preserves an unimpeded identity path while allowing each block to learn a bounded residual adjustment G_l, which mitigates vanishing gradients and supports deeper compositions without relying on any particular normalization or activation choice. Subsequent to the composition of the L blocks, a linear head produces the descriptor embedding:

$$z = W^{(h)} norm\big(h^{(L)}\big) + b^{(h)} \tag{8}$$

which serves as the final representation for downstream tasks. The resulting vector z serves as the descriptor-based embedding. We initialize linear weights with a Kaiming-normal scheme consistent with the chosen nonlinearity [31].

2.3.4 Fusion layer

We fuse multi-modal features using computationally efficient dimension-wise attention across modalities. The fusion module learns per-dimension attention weights normalized across modalities via softmax. A scalar gate parameter controls interpolation between learned attention and uniform averaging. This produces a weighted combination of modality-specific features for each dimension. Because weights are dimension-specific, different dimensions can emphasize different modalities. The fused vector is determined as follows, given the feature vector *x* from modality *m* and optional bias *b*:

$$y = \sum_{m=1}^{M} \alpha_m x_m + b \tag{9}$$

Where each a_m being a non-negative per-dimension weight vector and the weights sum to one across modalities for each dimension. The weights a_m being obtained by applying a softmax to learned modality scores and then optionally interpolating with the uniform average via a scalar gate *g*. In the case of $g = 0$ relies entirely on learned attention, while $g = 1$ reduces to a simple average across modalities. Missing modalities are replaced with zero vectors, allowing the attention mechanism to dynamically down-weight their contribution. The resulting fused vector is passed to species-specific independent networks.

2.3.5 Prediction

Each species-specific prediction is generated by independent networks applied to the fused representation. Each head is a three-layer MLP. The first layer projects the fused vector to dimension with ReLU activation and dropout. The second layer reduces to dimension, and the final layer outputs a scalar prediction for binary classification. This design enables transfer learning through shared representations while maintaining independent networks.

2.3.6 Evaluation and hyperparameter selection

We employed stratified 10-fold cross-validation to maintain balanced class distributions across folds [32]. We used the AdamW optimizer with learning rate 1e-4 and weight decay 1e-4 [33]. Models were trained for up to 200 epochs with early stopping (patience = 10) based on validation loss. The checkpoint with the lowest validation loss was used for testing. We applied cosine annealing to the learning rate (T_max=200, η_min=1e-6) [34]. We optimized binary cross-entropy loss with logits. Model performance was evaluated using AUROC and AUPRC. A threshold of 0.5 was used to generate binary predictions [35].

2.3.7 Shap and EdgeSHAPer for model interpretation

We employed SHAP (SHapley Additive exPlanations) [6] to quantify the contribution of ADME and physicochemical descriptors to each prediction, revealing both the magnitude and direction of each descriptor's influence [36]. For each species, we fixed fingerprint and graph features at their background values and varied only the ADME descriptor vector. We aggregated individual sample attributions across the test set and ranked features by mean absolute SHAP value.

Next, we applied EdgeSHAPer [7] to attribute each prediction to specific bonds. Adjacent bond contributions were aggregated into molecular fragments. For each fragment, we computed the mean contribution weighted by fragment frequency in the test set. Positive values indicate stabilizing effects; negative values indicate destabilizing effects. We then analyzed fragment contributions and their consistency across species.

To assess fragment-ADME associations, we partitioned molecules by fragment presence/absence and tested for enrichment in ADME properties. For continuous ADME properties, we quantified the effect size using Cohen's d and tested significance with the Mann-Whitney U test [37]. For binary ADME properties, we computed log-odds ratios and assessed significance using Fisher's exact test [38]. We applied the Benjamini-Hochberg procedure to control the false discovery rate across multiple comparisons. Results are visualized as heatmaps where color indicates effect direction and magnitude, with markers denoting FDR-significant associations.

3. Results

3.1 Exploratory Data Analysis

An exploratory data analysis (EDA) was conducted to assess data quality, split comparability, and chemical-space coverage prior to model training. The training and test sets were first compared in terms of fundamental molecular properties to ensure similar chemical space coverage. Inappropriate train-test splits, where test compounds systematically differ from training examples in physicochemical properties, can lead to artificially optimistic or pessimistic performance estimates that fail to reflect real-world model utility [39]. Scatter plots of molecular weight (MW) against LogP, MW against TPSA, and LogP against TPSA for training and test compounds (Fig. 1A-C) showed largely overlapping distributions. The test set spanned a range of physicochemical properties comparable to the training set, with no evidence of systematic shifts between datasets. A small number of compounds (<5%) with extreme properties (e.g., MW >1000 Da, LogP >10, or TPSA >300 Å^2) were present in both sets, likely reflecting specialized compound classes such as macrocycles or highly conjugated systems. This indicates that the test compounds occupy a comparable chemical domain to the training compounds.

To further assess structural novelty and diversity, we computed nearest-neighbor similarity metrics and diversity indices for each subset. Quantifying chemical diversity is critical, as models trained on narrowly clustered chemical series often overfit and fail to generalize to new scaffolds [40]. The mean and median k-nearest-neighbor (k-NN) Tanimoto similarities of test compounds to the training set were notably lower than those of training compounds among themselves (Fig. 1D). Corresponding diversity indices were 0.36 for the training set and 0.53 for the test set, indicating that test compounds are structurally more diverse on average. To complement these summary statistics, we analyzed the cumulative distribution of nearest-neighbor similarities (Fig. 1E). The training set curve rose steeply at low similarity values (< 0.2) and plateaued rapidly, indicating that the vast majority (> 90%) of training compounds had at least one highly similar neighbor (Tanimoto > 0.5) within the training pool. In contrast, the cumulative curve for the test set increased more gradually across the entire similarity range. Notably, 43% of test compounds exhibited maximum Tanimoto similarities below 0.5 to their nearest training neighbor, compared to only 3% among training compounds. This distribution difference confirms that while the training and test sets occupy overlapping chemical space in terms of global physicochemical properties, the test set includes structurally novel scaffolds that are sufficiently distinct from training examples.

Multi-task learning has been shown to improve predictive performance by leveraging shared structure-response relationships across related endpoints, but its utility depends on meaningful inter-task correlation [16]. A heatmap of pairwise Phi coefficients revealed moderate positive correlations in liver microsomal stability across human, mouse, and rat species (Fig. 1F). Specifically, the Phi values were 0.58 for human-mouse, 0.58 for human-rat, and 0.63 for mouse-rat comparisons. This partial alignment suggests that while species-specific metabolic processes exist reflecting differences in cytochrome P450 isoform expression and substrate preferences there is a consistent underlying signal that can be exploited via joint modeling [41]. The presence of cross-species correlation justifies the application of a multi-task framework to leverage shared metabolic determinants while preserving species-specific distinctions.

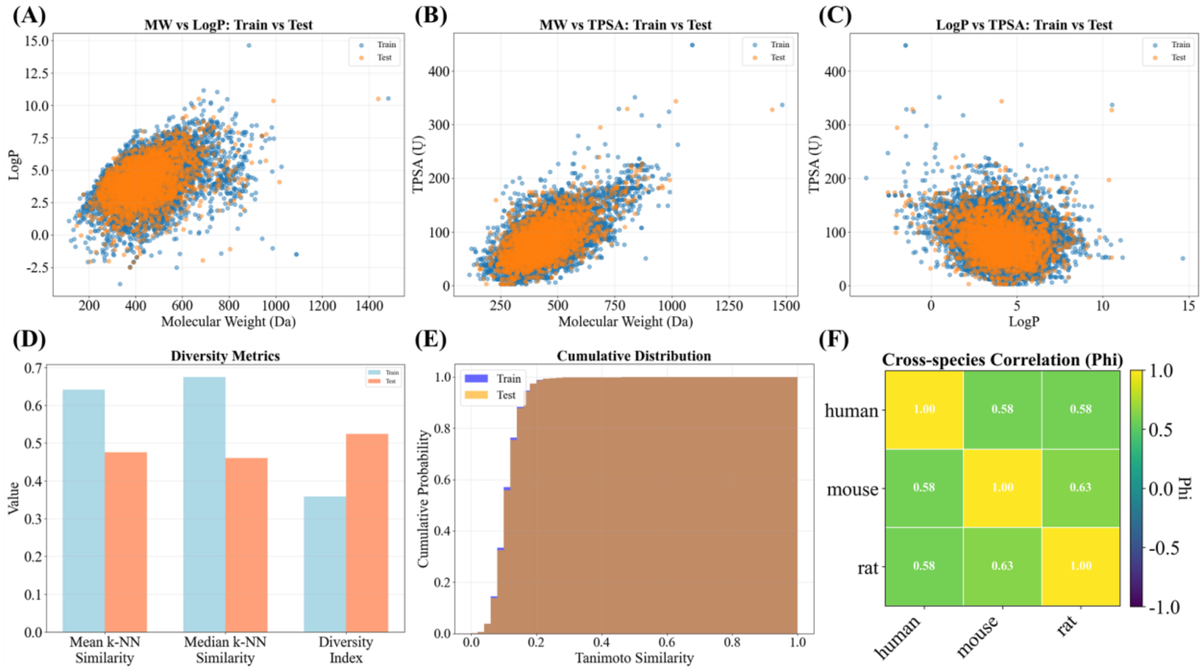

Fig. 3. Exploratory data analysis of liver microsomal stability dataset and cross-species label concordance. Overlaid scatter plots for key physicochemical pairs comparing Train (blue) and Test (orange): (A) molecular weight (MW) plotted against logP, (B) MW plotted against topological polar surface area (TPSA), and (C) logP plotted against TPSA. Distributions show substantial visual overlap between splits. (D) Mean and median k-NN Tanimoto similarity and a diversity index. (E) Empirical cumulative distribution functions (ECDFs) of k-NN Tanimoto similarity for Train and Test, providing a bin-free view of nearest-neighbor similarity. (F) Cross-species correlation of binary labels summarized by the Phi coefficient.

3.2 Performance evaluation

To evaluate our model's classification performance, we compared three modeling strategies using stratified 10-fold cross-validation. These included single-task models trained separately for each species, single-modal models using SMILES features only, and our full multi-task, multi-modal model. Fig. 4 shows AUROC and AUPRC, averaged across folds with 95% confidence bands. In HLM, our model achieves the highest AUPRC (0.712 ± 0.001) and ties the best AUROC (0.770 ± 0.001). This outperforms the single-modal (0.705 ± 0.001) and single-task (0.696 ± 0.001) baselines on AUPRC. In RLM, the multi-task, multi-modal approach shows the most substantial improvement. Our model achieves the top scores (0.785 ± 0.001 and 0.715 ± 0.001), exceeding single-modal (0.777 ± 0.001 and 0.703 ± 0.001) and single-task (0.762 ± 0.001 and 0.706 ± 0.001). For MLM, performance is comparable between approaches. The single-task model slightly leads on AUROC and AUPRC (0.774 ± 0.001 and 0.691 ± 0.001), while our model achieves comparable performance (0.766 ± 0.001 and 0.679 ± 0.001), suggesting that threshold selection can be adjusted based on application requirements. Across all species, the single-modal baseline consistently underperforms relative to multi-modal approaches, indicating that ADME and physicochemical features provide complementary information to structural encodings. Overall, multi-task, multi-modal learning improves cross-species generalization and yields the most balanced performance for HLM and RLM while remaining competitive for MLM. Additional metrics (precision, recall, F1-score, specificity) are provided in Appendix Table S1.

Table 1. Cross-species performance under alternative modeling settings. A set of evaluation metrics was employed to assess the performance of both models, including AUROC, AUPR.

Model	Species	AUROC	AUPR
Single task	Human	0.770	0.696
	Rat	0.778	0.703
	Mouse	**0.774**	**0.691**
Single modal (only smiles)	Human	0.761	0.705
	Rat	0.775	0.706
	Mouse	0.739	0.649
Our model	Human	**0.770**	**0.712**
	Rat	**0.785**	**0.715**
	Mouse	0.766	0.679

3.3 SHAP-based interpretation of ADME determinants of liver microsomal stability across species

We conducted a transparent SHAP analysis of the liver microsomal stability model (Fig. 5). Throughout this study, positive model output (and positive SHAP values) denotes the “unstable” class ($t_{1/2} \leq 30$ min). SHAP, grounded in cooperative game theory, quantifies each feature’s contribution to the model output and enables sample-level interpretation.

As shown in the beeswarm plots (Fig. 5A-C), multiple descriptors exhibit consistently dense clustering across HLM, RLM, and MLM tasks. These include enzyme-interaction flags (CYP3A4, CYP2D6, CYP2C19 inhibitors), membrane permeability and transport indicators (BBB-permeant, P-gp substrate), descriptors along the lipophilicity-polarity axis (logP-family metrics: iLOGP, MLOGP, XLOGP3, WLOGP; TPSA; hydrogen-bond donors/acceptors) [42], and rule-based indices (Ghose, Egan, Veber, Lipinski violations). Examining both point color (feature value) and horizontal position (SHAP value sign), higher lipophilicity (logP family) shifts predictions toward "unstable," whereas higher polarity (TPSA, H-bond counts) shifts them toward "stable" [43]. The binary features BBB-permeant and P-gp substrate show clear bimodal SHAP distributions, indicating that membrane permeability and transport propensity systematically influence predictions [44].

The mean absolute SHAP bar plots (Fig. 5D-F) provide quantitative support for these observations. Enzyme-inhibition flags, permeability indicators, and lipophilicity-polarity descriptors consistently rank among the top ten contributors across all three species. Notably, the top three to five descriptors exhibit significantly larger mean absolute SHAP values than the remaining features, indicating that a small subset of ADME descriptors accounts for most of the model's explanatory power.

Species-specific differences are evident [45, 46]. In HLM, CYP3A4, CYP2D6, and CYP2C19 inhibitor flags, BBB-permeant, P-gp substrate, and logP-family metrics rank among the top contributors, alongside physicochemical properties such as molar refractivity (MR) (Fig. 5A,D) [47, 48]. In RLM, BBB-permeant and Ghose violations are comparatively prominent, with the logP-TPSA axis also ranking highly (Fig. 5B, E). In MLM, CYP2C19 and CYP3A4 inhibitor flags exert a particularly pronounced influence (Fig. 5C, F). These are learned correlations but align with known causal mechanisms in microsomal metabolism.

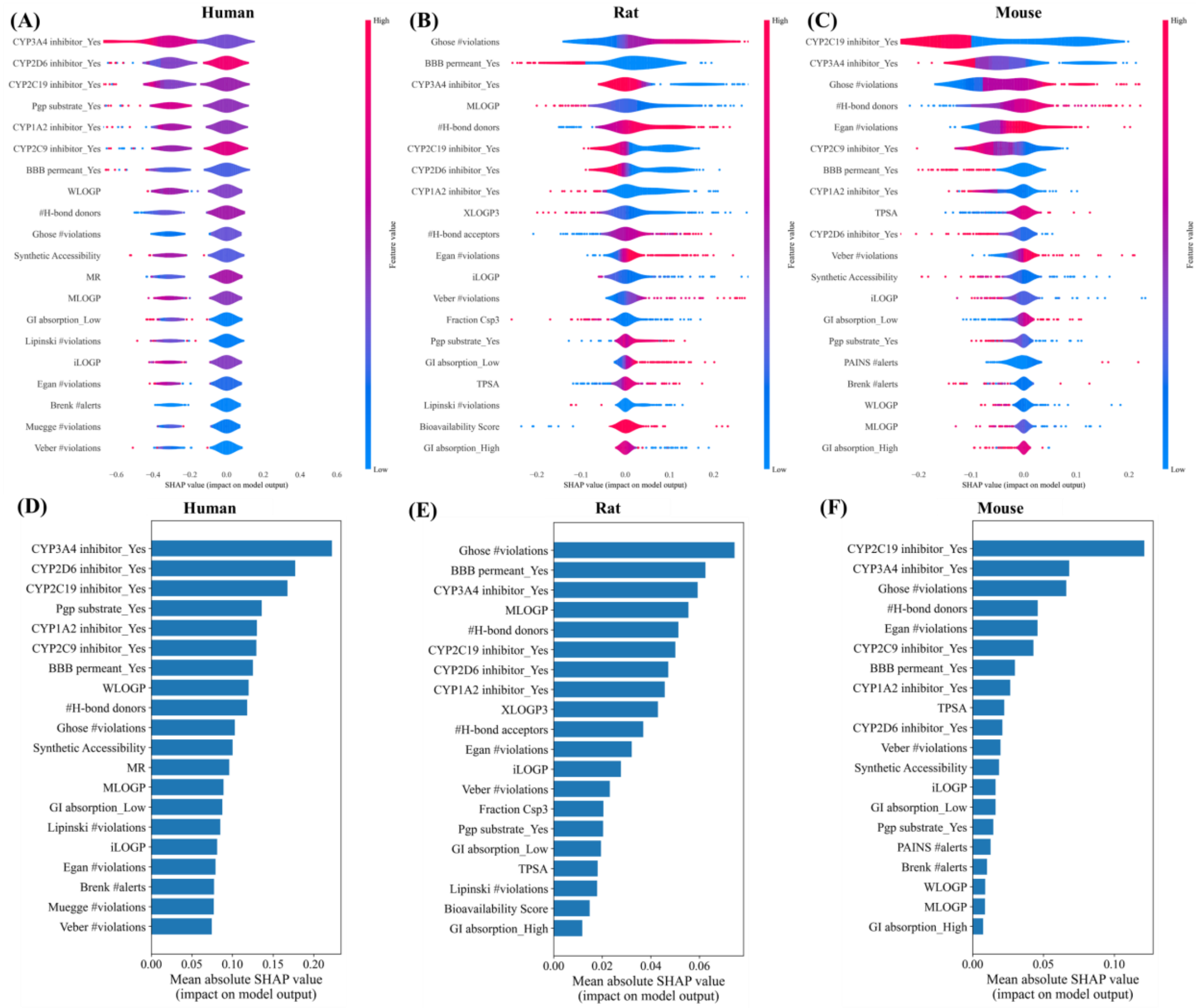

Fig. 5. SHAP-based interpretation of ADME descriptors across species Beeswarm plots showing SHAP value distributions for molecular descriptors in (A) human (HLM), (B) rat (RLM), and (C) mouse (MLM) liver microsomal stability predictions. Each point represents a sample, with color indicating feature value (red: high, blue: low) and horizontal position showing SHAP value (contribution to model output). Bar plots of the top 10 most important descriptors ranked by mean absolute SHAP value for (D) HLM, (E) RLM, and (F) MLM stability predictions.

3.4 Fragment-level explanations of metabolic stability per task

This analysis uses EdgeSHAPer-derived edge-importance scores to quantify each (Cohen's d) on the held-out test set. In the visualizations, red fragments (Fig. 6A, B, C) indicate destabilizing features associated with decreased microsomal stability, whereas blue fragments (Fig. 6D, E, F) indicate stabilizing features. Localizing these zones within individual molecules provides actionable guidance for structural optimization.

Several bond-level patterns recur across HLM, RLM, and MLM. Alkenes and allylic/benzylic environments act as destabilizing metabolic features, prominently represented among the top fragments in HLM (Fig. 6A), RLM (Fig. 6B), and MLM (Fig. 6C) [49, 50]. Amide- and carbamate-containing carbonyl motifs have been observed to appear on the stabilizing side under oxidative microsomal conditions, featuring among the leading fragments in HLM (Fig. 6D) and likewise in RLM and MLM (Fig. 6E, F) [51, 52]. Conversely, nitriles, halogens, and high-oxidation-state sulfur groups exhibit context- and species-dependent behavior. Specifically, a nitrile-bearing fragment is observed on the stabilizing list for RLM (Fig. 6E), while a simple nitrile fragment contributes to instability in MLM (Fig. 6F) [53]. Multiple halogenated fragments are identified among destabilizing features in both RLM and MLM (Fig. 6B, C) [54]. Notably, a sulfonamide-type S(VI) fragment is among the most destabilizing features in MLM (Fig. 6E) [55].

Under NADPH-dependent microsomal oxidation, recurrent fragment-outcome relationships emerge across HLM, RLM, and MLM. Alkenes and allylic/benzylic carbons act as robust destabilizing features [49], consistent with cytochrome P450-mediated allylic C-H hydroxylation and C=C epoxidation that preferentially target these motifs and often associated with clearance (Fig. 6A, C, E) [56, 57]. By contrast, nitrile groups (-C≡N) display pronounced scaffold- and species-dependence: a vinyl-nitrile context associates with stabilization in RLM (Fig. 6E), whereas a simple CN fragment contributes to destabilization in MLM (Fig. 6F) [58, 59]. These paired observations indicate that neighboring functionality and species-specific metabolism can override the anticipated metabolic-masking effect. Sulfur-containing functionalities show oxidation-state dependent behavior, with thioethers and thiocarbonyls appearing on both sides across species consistent with stepwise oxidation to sulfoxides/sulfones and, in certain contexts, the formation of reactive intermediates. In keeping with this propensity, an S(VI) sulfonamide motif is among the destabilizing features in MLM (Fig. 6C) [60]. Halogenation, although commonly used to block oxidative soft spots, shows a nuanced, position- and species-dependent role; several halogenated fragments rank among the top destabilizing motifs in RLM and MLM (Fig. 6B, C) [61]. Finally, carbonyl-containing motifs tend to correlate with stabilization under NADPH-dependent oxidative conditions (Fig. 6D, E, F) [51], while acknowledging that scaffold-specific hydrolytic pathways can predominate for particular chemotypes. Collectively, these patterns support a context- and species-dependent view of oxidative propensity and mitigation, underscoring the value of fragment-level interpretation for hypothesis-driven structural optimization.

In sum, the cross-species data substantiates fundamental metabolic principles while refining their practical application in design. Alkene and allylic features have been shown to reproducibly destabilize (Fig. 6A, B, C), and a significant number of amide-rich frameworks have been found to be comparatively stable under oxidative microsomal conditions (Fig. 6D, E, F). However, nitriles, halogens, and even S(VI) sulfur groups exhibit marked dependency on local structure and species (Fig. 6C, D), highlighting the value of fragment-level attribution for prioritizing modifications that enhance microsomal stability.

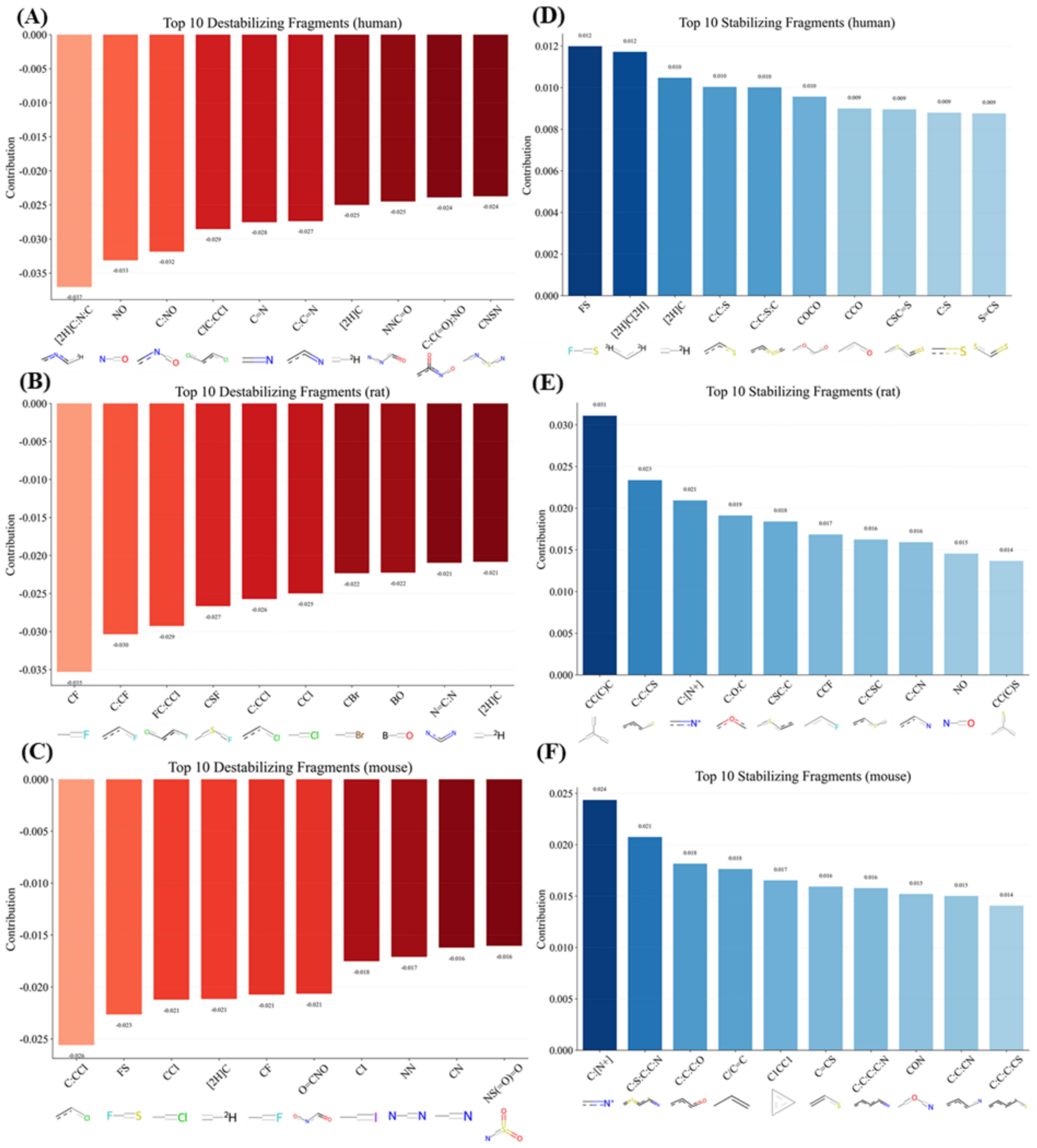

Fig. 6. Fragment-level explanations of microsomal stability across species. Top 10 molecular fragments with the greatest impact on metabolic stability predictions. Bar plots showing bond-level SHAP contributions for destabilizing fragments (negative values, red bars) in (A) human (HLM), (B) rat (RLM), and (C) mouse (MLM), and stabilizing fragments (positive values, blue bars) in (D) HLM, (E) RLM, and (F) MLM liver microsomal stability. Representative molecular structures with highlighted bonds are shown below each bar.

3.5 Fragment-ADME enrichment across HLM, RLM, and MLM

Using the top 10 fragments highlighted by EdgeSHAPer, we mapped fragment-ADME associations across HLM, RLM, and MLM and observed consistent medium-to-large effect sizes (Fig. 7A-F). Halogenated fragments, especially C-Cl and CF_3-containing motifs, were associated with higher lipophilicity (Consensus logP, XLOGP3, WLOGP), greater molar refractivity, and lower predicted aqueous solubility (ESOL logS) (Fig. 7A-C) [62]. In several instances, these fragments were also enriched among compounds predicted as BBB-permeant (Fig. 7D-F) [63]. This pattern is chemically coherent: halogenation particularly CF_3 substitution typically increases lipophilicity and molar refractivity and can attenuate oxidative lability when substitution blocks oxidation-prone C-H positions [64, 65]. The net impact on microsomal stability is context-dependent. However, when overall polarity is approximately constant, increased lipophilicity combined with reduced oxidative susceptibility leads to greater passive membrane transport and, for some scaffolds, higher brain penetration. This aligns with the observed enrichment signal (Fig. 7D-F) [63, 64].

Fragments bearing amide or carbamate functionality showed the opposite trend, with lower logP, higher ESOL logS, and reduced predicted BBB-permeation [66]. This is consistent with established BBB heuristics, which indicate that passive CNS exposure declines as topological polar surface area increases, particularly beyond about 60-90 Å^2. The association also accords with the ESOL model, where logP, molecular weight, aromaticity, and rotatable bond count are key predictors of solubility [67]. In addition to the polarity-permeability trade-off,, several fragments previously identified as destabilizing aligned with canonical CYP-mediated oxidation motifs, most notably benzylic and allylic C-H bonds (Fig. 7D-F) [49, 68]. These well-characterized metabolic hot spots provide a mechanistic basis for the unfavorable stability associations. In contrast, saturated or branched aliphatic fragments displayed ADME profiles that differed from flat, highly aromatic motifs, in line with the escape-from-flatland observation that increasing sp^3 content can improve developability (Fig. 7A-C) [69, 70]. Finally, fragments with greater conformational flexibility and amphipathicity co-occurred with enrichment for P-gp substrates (Fig. 7D-F) [71], which is consistent with reports that P-glycoprotein preferentially recognizes hydrophobic, amphipathic scaffolds and actively exports them.

We interpreted effect sizes using pre-specified thresholds. For continuous endpoints, we highlighted fragments with the magnitude of Cohen’s d in the moderate-to-large range (Fig. 7A-C) [72]. For binary endpoints, we emphasized enrichments with $|\log(OR)| \geq 0.69$, which corresponds to $OR \geq 2$ (Fig. 7D-F) [73]. The heatmaps are intended to aid pattern recognition and should be read as complements to the tables rather than substitutes [74]. Modest species-level differences are consistent with interspecies variation in hepatic enzyme expression across human, rat, and mouse [41].

Taken together, the enrichment landscape yields a coherent set of design signals. Halogenation, particularly CF_3 placement near oxidation-susceptible positions, increases lipophilicity and can enhance stability by blocking vulnerable C-H sites and attenuating oxidative metabolism [75]. This benefit typically entails a predictable solubility penalty and, when polarity is held approximately constant, a greater likelihood of BBB permeation [76]. In contrast, carbonyl- and heteroatom-rich fragments tend to increase solubility and reduce BBB propensity at the expense of permeability, consistent with PSA-based CNS heuristics and the ESOL framework [63]. Aromatic, allylic, and benzylic motifs are indicative of metabolic propensity, in line with established p450 chemistry (Fig. 7D-F) [49]. Increasing sp^3 character offers a practical countermeasure to these destabilizing features and can improve developability

(Fig. 7A-C) [69]. Across species, these trends show limited divergence and provide actionable hypotheses for fragment substitution and placement when optimizing microsomal stability alongside broader ADME properties [77].

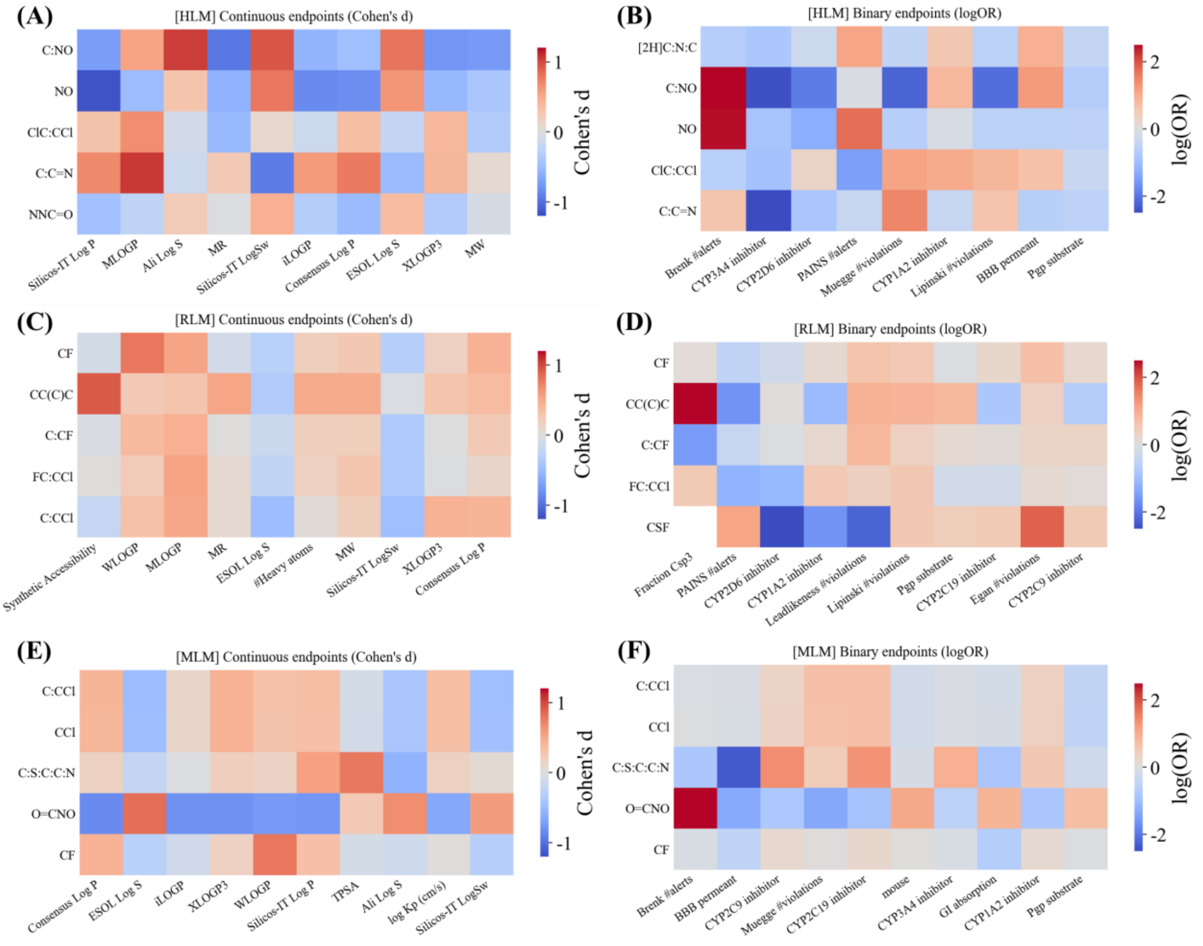

Fig. 7. Fragment-ADME enrichment across species. Heatmaps showing the association between top metabolic stability fragments and various ADME/physicochemical properties. Effect sizes for continuous endpoints (Cohen's d) are displayed for (A) HLM, (B) RLM, and (C) MLM, while effect sizes for binary endpoints (log odds ratio) are shown for (D) HLM, (E) RLM, and (F) MLM. Rows represent the top 10 EdgeSHAPer fragments identified for each species, and columns represent different ADME properties or molecular classifications.

4. Discussion

This study demonstrates the effectiveness of a cross-species multi-task learning framework that integrates multi-modal molecular representations for predicting liver microsomal stability. Cross-validated performance shows consistent gains over single-modal and single-task baselines in HLM and RLM, and competitive performance in MLM, aligning with the design intent: a shared representation learns general structure-metabolism patterns, while task-specific heads capture idiosyncratic species effects. Descriptor-level SHAP analysis revealed that permeability/transport indicators, enzyme-interaction flags, and the lipophilicity-polarity axis strongly influence predictions. Higher lipophilicity shifted predictions toward instability, while greater polarity shifted them toward stability. Binary transport features established clear decision boundaries. Graph-level EdgeSHAPer localized bond- and fragment-level effects: alkenes and allylic/benzylic contexts frequently acted as destabilizing features, while amide/carbamate carbonyl motifs often conferred stability. Nitriles, halogens, and higher-oxidation-state sulfur groups showed context-dependent, species-variable contributions. Fragment-ADME enrichment analysis showed that halogenated motifs exhibited higher logP and molar refractivity, lower solubility, and greater propensity for BBB permeation, while amide/carbamate-rich fragments showed the opposite trend. Increased saturation and sp^3 character were associated with developability-favored profiles. These patterns suggest practical design strategies: tuning lipophilicity/polarity, modulating flexibility, and shielding labile positions. Beyond these chemical insights, this work also makes methodological contributions by integrating complementary molecular representations within a unified learning framework and combines global and local interpretability. This approach advances from opaque screening to hypothesis-driven optimization and offers a principled basis for prioritizing synthesis. However, several limitations should be addressed. First, labels compiled from heterogeneous PubChem BioAssay sources were binarized at a fixed half-life threshold, which compresses kinetic information and can introduce assay-specific variability. To mitigate this, future work should model continuous or ordinal half-life values to retain kinetic information. Additionally, incorporating assay identifiers and covariates with random effects can account for assay-level variance, while sensitivity analyses across thresholds can assess robustness. Second, cross-validation indicates stable internal generalization but does not guarantee robustness under distribution shift across laboratories or across species. To address this, future work should evaluate performance on external test sets, employ scaffold-based and temporal splits to assess generalization, and explore domain adaptation methods to improve cross-species portability. Third, *in silico* ADME descriptors inherit assumptions and potential biases from their source models, and class imbalance complicates the choice of operating points and calibration. Potential mitigations include augmenting descriptors with experimental measurements where available, applying probability calibration methods such as Platt scaling, quantifying uncertainty via ensemble approaches, and addressing class imbalance through cost-sensitive learning or focal loss.

5. Conclusions

This study demonstrates that a multi-modal, multi-task learning framework effectively predicts liver microsomal stability across HLM, RLM, and MLM with chemically coherent explanations. Descriptor-level SHAP emphasizes transport/permeability features, enzyme-interaction flags, and lipophilicity-polarity balance. EdgeSHAPer localizes stabilizing and destabilizing substructures consistent with known metabolic destabilizing features and protective motifs. Fragment-ADME enrichment provides actionable guidance for structural optimization: adjusting lipophilicity and polarity, increasing sp^3 content, blocking labile positions, and preserving stabilizing carbonyl motifs. Collectively, these findings demonstrate that integrating structural and ADME representations within a cross-species framework enhances predictive performance and facilitates rational design to mitigate microsomal instability.

CRediT authorship contribution statement

Subhin Seomun: Writing - original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sunyong Yoo**: Writing - review & editing, Supervision, Resources, Project administration, Funding acquisition.

Data availability

The implementation of the proposed model and the preprocessed data are available at: https://github.com/bmil-jnu/ADME-enhanced_multitask_prediction_of_microsomal_stability.git

Declaration of competing interest

The authors declare the existence of no competing interests.

Acknowledgment

This work was supported by the Ministry of Food and Drug Safety (MFDS) grants RS-2024-00332003 (2024) and RS-2025-02215961 (2025), the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare (MOHW), Republic of Korea, grant number RS-2025-19252970, and the Ministry of Science and ICT (MSIT) support program grant number RS-2025-16063391.

References

[1] V.B. Siramshetty, P. Shah, E. Kerns, K. Nguyen, K.R. Yu, M. Kabir, J. Williams, J. Neyra, N. Southall, Đ.-T. Nguyễn, Retrospective assessment of rat liver microsomal stability at NCATS: data and QSAR models, Scientific reports, 10 (2020) 20713.

[2] K. Słoczyńska, A. Gunia-Krzyżak, P. Koczurkiewicz, K. Wójcik-Pszczoła, D. Żelaszczyk, J. Popiół, E. Pękala, Metabolic stability and its role in the discovery of new chemical entities, Acta Pharmaceutica, 69 (2019) 345-361.

[3] M. Varma, S. Steyn, C. Allerton, A. El-Kattan, Predicting Clearance Mechanism in Drug Discovery: Extended Clearance Classification System (ECCS), Pharmaceutical research, 32 (2015).

[4] G. Camenisch, Drug Disposition Classification Systems in Discovery and Development: A Comparative Review of the BDDCS, ECCS and ECCCS Concepts, Pharmaceutical research, 33 (2016).

[5] Z. Cheng, X. Zhou, Z. Du, W. Li, B. Hu, J. Tian, L. Zhang, J. Huang, H. Jiang, Metabolic stability and metabolite characterization of capilliposide B and capilliposide C by LC–QTRAP–MS/MS, Pharmaceutics, 10 (2018) 178.

[6] A. Wojtuch, R. Jankowski, S. Podlewska, How can SHAP values help to shape metabolic stability of chemical compounds?, Journal of cheminformatics, 13 (2021) 74.

[7] A. Mastropietro, G. Pasculli, C. Feldmann, R. Rodríguez-Pérez, J. Bajorath, EdgeSHAPer: Bond-centric Shapley value-based explanation method for graph neural networks, Iscience, 25 (2022).

[8] L. Di, E.H. Kerns, Y. Hong, T.A. Kleintop, O.J. Mc Connell, D.M. Huryn, Optimization of a higher throughput microsomal stability screening assay for profiling drug discovery candidates, SLAS Discovery, 8 (2003) 453-462.

[9] F.P. Guengerich, Cytochrome p450 and chemical toxicology, Chemical research in toxicology, 21 (2008) 70-83.

[10] S. Podlewska, R. Kafel, MetStabOn—online platform for metabolic stability predictions, International journal of molecular sciences, 19 (2018) 1040.

[11] J.Y. Ryu, J.H. Lee, B.H. Lee, J.S. Song, S. Ahn, K.-S. Oh, PredMS: a random forest model for predicting metabolic stability of drug candidates in human liver microsomes, Bioinformatics, 38 (2022) 364-368.

[12] B.-X. Du, Y. Long, X. Li, M. Wu, J.-Y. Shi, CMMS-GCL: cross-modality metabolic stability prediction with graph contrastive learning, Bioinformatics, 39 (2023) btad503.

[13] T.-Z. Long, D.-J. Jiang, S.-H. Shi, Y.-C. Deng, W.-X. Wang, D.-S. Cao, Enhancing Multi-species Liver Microsomal Stability Prediction through Artificial Intelligence, Journal of Chemical Information and Modeling, 64 (2024) 3222-3236.

[14] T. Wang, Z. Li, L. Zhuo, Y. Chen, X. Fu, Q. Zou, MS-BACL: enhancing metabolic stability prediction through bond graph augmentation and contrastive learning, Briefings in Bioinformatics, 25 (2024) bbae127.

[15] A. Lamens, J. Bajorath, Explaining multiclass compound activity predictions using counterfactuals and shapley values, Molecules, 28 (2023) 5601.

[16] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, V. Pande, Massively multitask networks for drug discovery, arXiv preprint arXiv:1502.02072, (2015).

[17] P. Shah, V.B. Siramshetty, A.V. Zakharov, N.T. Southall, X. Xu, D.-T. Nguyen, Predicting liver cytosol stability of small molecules, Journal of cheminformatics, 12 (2020) 21.

[18] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, PubChem 2025 update, Nucleic acids research, 53 (2025) D1516-D1525.

[19] M.A. Zientek, K. Youdim, Reaction phenotyping: advances in the experimental strategies used to characterize the contribution of drug-metabolizing enzymes, Drug Metabolism and Disposition, 43 (2015) 163-181.

[20] J. Lee, J.L. Beers, R.M. Geffert, K.D. Jackson, A review of CYP-mediated drug interactions: mechanisms and in vitro drug-drug interaction assessment, Biomolecules, 14 (2024) 99.

[21] P. Shah, V.B. Siramshetty, E. Mathé, X. Xu, Developing robust human liver microsomal stability prediction models: Leveraging inter-species correlation with rat data, Pharmaceutics, 16 (2024) 1257.

[22] V.D. Hähnke, S. Kim, E.E. Bolton, PubChem chemical structure standardization, Journal of cheminformatics, 10 (2018) 36.

[23] D. Rogers, M. Hahn, Extended-connectivity fingerprints, Journal of chemical information and modeling, 50 (2010) 742-754.

[24] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL keys for use in drug discovery, Journal of chemical information and computer sciences, 42 (2002) 1273-1280.

[25] G. Landrum, Rdkit documentation, Release, 1 (2013) 4.

[26] L. Xie, L. Xu, R. Kong, S. Chang, X. Xu, Improvement of prediction performance with conjoint molecular fingerprint in deep learning, Frontiers in pharmacology, 11 (2020) 606668.

[27] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, Advances in neural information processing systems, 28 (2015).

[28] B. Jiang, Z. Zhang, D. Lin, J. Tang, B. Luo, Semi-supervised learning with graph learning-convolutional networks, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 11313-11320.

[29] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Scientific reports, 7 (2017) 42717.

[30] Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko, Revisiting deep learning models for tabular data, Advances in neural information processing systems, 34 (2021) 18932-18943.

[31] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026-1034.

[32] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Ijcai, Montreal, Canada, 1995, pp. 1137-1145.

[33] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101, (2017).

[34] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983, (2016).

[35] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PloS one, 10 (2015) e0118432.

[36] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems, 30 (2017).

[37] J. Cohen, Statistical power analysis for the behavioral sciences, routledge2013.

[38] A. Agresti, M. Kateri, Categorical data analysis, International Encyclopedia of Statistical Science, Springer2025, pp. 408-411.

[39] R.P. Sheridan, Time-split cross-validation as a method for estimating the goodness of prospective prediction, Journal of chemical information and modeling, 53 (2013) 783-790.

[40] E.J. Martin, V.R. Polyakov, X.-W. Zhu, L. Tian, P. Mukherjee, X. Liu, All-assay-Max2 pQSAR: activity predictions as accurate as four-concentration IC50s for 8558 Novartis assays, Journal of chemical information and modeling, 59 (2019) 4450-4459.

[41] M. Martignoni, G.M. Groothuis, R. de Kanter, Species differences between mouse, rat, dog, monkey and human CYP-mediated drug metabolism, inhibition and induction, Expert opinion on drug metabolism & toxicology, 2 (2006) 875-894.

[42] A. Daina, O. Michielin, V. Zoete, iLOGP: a simple, robust, and efficient description of n-octanol/water partition coefficient for drug design using the GB/SA approach, Journal of chemical information and modeling, 54 (2014) 3284-3301.

[43] D.F. Veber, S.R. Johnson, H.-Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, Journal of medicinal chemistry, 45 (2002)

2615-2623.

[44] A. Schinkel, J. Smit, m. van Tellingen, J. Beijnen, E. Wagenaar, L. Van Deemter, C. Mol, M. Van der Valk, E. Robanus-Maandag, H. Te Riele, Disruption of the mouse mdr1a P-glycoprotein gene leads to a deficiency in the blood-brain barrier and to increased sensitivity to drugs, Cell, 77 (1994) 491-502.

[45] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Advanced drug delivery reviews, 23 (1997) 3-25.

[46] D.J. Huggins, A.R. Venkitaraman, D.R. Spring, Rational methods for the selection of diverse screening compounds, ACS chemical biology, 6 (2011) 208-217.

[47] S.A. Wildman, G.M. Crippen, Prediction of physicochemical parameters by atomic contributions, Journal of chemical information and computer sciences, 39 (1999) 868-873.

[48] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases, Journal of combinatorial chemistry, 1 (1999) 55-68.

[49] P.R. Ortiz de Montellano, Hydrocarbon hydroxylation by cytochrome P450 enzymes, Chemical reviews, 110 (2010) 932-948.

[50] Z. Zhang, W. Tang, Drug metabolism in drug discovery and development, Acta Pharmaceutica Sinica B, 8 (2018) 721-732.

[51] M. Lang, U.S. Ganapathy, L. Mann, R.W. Seidel, R. Goddard, F. Erdmann, T. Dick, A. Richter, Synthesis and in vitro metabolic stability of sterically shielded antimycobacterial phenylalanine amides, ChemMedChem, 19 (2024) e202300593.

[52] Y. Liu, C. Ma, Y. Li, M. Li, T. Cui, X. Zhao, Z. Li, H. Jia, H. Wang, X. Xiu, Design, synthesis and biological evaluation of carbamate derivatives incorporating multifunctional carrier scaffolds as pseudo-irreversible cholinesterase inhibitors for the treatment of Alzheimer's disease, European Journal of Medicinal Chemistry, 265 (2024) 116071.

[53] S.P. Rendic, F.P. Guengerich, Formation of potentially toxic metabolites of drugs in reactions catalyzed by human drug-metabolizing enzymes, Archives of toxicology, 98 (2024) 1581-1628.

[54] J.B. Behrendorff, Reductive cytochrome P450 reactions and their potential role in bioremediation, Frontiers in Microbiology, 12 (2021) 649273.

[55] S.C. Khojasteh, U.A. Argikar, J.P. Driscoll, C.J. Heck, L. King, K.D. Jackson, W. Jian, A.S. Kalgutkar, G.P. Miller, V. Kramlinger, Novel advances in biotransformation and bioactivation research–2020 year in review, Drug metabolism reviews, 53 (2021) 384-433.

[56] E.M. Isin, F.P. Guengerich, Complex reactions catalyzed by cytochrome P450 enzymes, Biochimica et Biophysica Acta (BBA)-General Subjects, 1770 (2007) 314-329.

[57] F.P. Guengerich, Cytochrome P450 oxidations in the generation of reactive electrophiles: epoxidation and related reactions, Archives of biochemistry and biophysics, 409 (2003) 59-71.

[58] F.F. Fleming, L. Yao, P. Ravikumar, L. Funk, B.C. Shook, Nitrile-containing pharmaceuticals: efficacious roles of the nitrile pharmacophore, Journal of medicinal chemistry, 53 (2010) 7902-7917.

[59] X. Wang, Y. Wang, X. Li, Z. Yu, C. Song, Y. Du, Nitrile-containing pharmaceuticals: target, mechanism of action, and their SAR studies, RSC medicinal chemistry, 12 (2021) 1650-1671.

[60] A.E. Cribb, S.P. Spielberg, G.P. Griffin, N4-hydroxylation of sulfamethoxazole by cytochrome P450 of the cytochrome P4502C subfamily and reduction of sulfamethoxazole hydroxylamine in human and rat hepatic microsomes, Drug metabolism and disposition, 23 (1995) 406-414.

[61] N.H. Cnubben, J. Vervoort, M.G. Boersma, I.M. Rietjens, The effect of varying halogen substituent patterns on the cytochrome P450 catalysed dehalogenation of 4-halogenated anilines to 4-aminophenol metabolites, Biochemical pharmacology, 49 (1995) 1235-1248.

[62] B. Jeffries, Z. Wang, R.I. Troup, A. Goupille, J.-Y. Le Questel, C. Fallan, J.S. Scott, E. Chiarparin, J. Graton, B. Linclau, Lipophilicity trends upon fluorination of isopropyl, cyclopropyl and 3-oxetanyl groups, Beilstein journal of organic chemistry, 16 (2020) 2141-2150.

[63] T.T. Wager, X. Hou, P.R. Verhoest, A. Villalobos, Central nervous system multiparameter optimization desirability: application in drug discovery, ACS chemical neuroscience, 7 (2016) 767-775.

[64] G. Chandra, D.V. Singh, G.K. Mahato, S. Patel, Fluorine-a small magic bullet atom in the drug development: perspective to FDA approved and COVID-19 recommended drugs, Chemical Papers, 77 (2023) 4085-4106.

[65] B. Jeffries, Z. Wang, J. Graton, S.D. Holland, T. Brind, R.D. Greenwood, J.-Y. Le Questel, J.S. Scott, E. Chiarparin, B. Linclau, Reducing the lipophilicity of perfluoroalkyl groups by CF2–F/CF2–Me or CF3/CH3 exchange, Journal of medicinal chemistry, 61 (2018) 10602-10618.

[66] J. Kelder, P.D. Grootenhuis, D.M. Bayada, L.P. Delbressine, J.-P. Ploemen, Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs, Pharmaceutical research, 16 (1999) 1514-1519.

[67] J.S. Delaney, ESOL: estimating aqueous solubility directly from molecular structure, Journal of chemical information and computer sciences, 44 (2004) 1000-1005.

[68] F.P. Guengerich, Mechanisms of cytochrome P450 substrate oxidation: MiniReview, Journal of biochemical and molecular toxicology, 21 (2007) 163-168.

[69] F. Lovering, J. Bikker, C. Humblet, Escape from flatland: increasing saturation as an approach to improving clinical success, Journal of medicinal chemistry, 52 (2009) 6752-6756.

[70] F. Lovering, Escape from Flatland 2: complexity and promiscuity, MedChemComm, 4 (2013) 515-519.

[71] A. Seelig, A general pattern for substrate recognition by P-glycoprotein, European Journal of Biochemistry, 251 (1998) 252-261.

[72] C.R. Brydges, Effect size guidelines, sample size calculations, and statistical power in gerontology, Innovation in aging, 3 (2019) igz036.

[73] S. Chinn, A simple method for converting an odds ratio to effect size for use in meta-analysis, Statistics in medicine, 19 (2000) 3127-3131.

[74] W.S. Cleveland, R. McGill, Graphical perception: Theory, experimentation, and application to the development of graphical methods, Journal of the American statistical association, 79 (1984) 531-

554.

[75] N.A. Meanwell, Fluorine and fluorinated motifs in the design and application of bioisosteres for drug design, Journal of medicinal chemistry, 61 (2018) 5822-5880.

[76] D.E. Clark, Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–brain barrier penetration, Journal of pharmaceutical sciences, 88 (1999) 815-821.

[77] I. Gardner, M. Xu, C. Han, Y. Wang, X. Jiao, M. Jamei, H. Khalidi, P. Kilford, S. Neuhoff, R. Southall, Non-specific binding of compounds in in vitro metabolism assays: a comparison of microsomal and hepatocyte binding in different species and an assessment of the accuracy of prediction models, Xenobiotica, 52 (2022) 943-956.

Appendix

Model	**Species**	**Recall**	**Specificity**	**Precision**	**F1-score**
Single task	Human	0.564	0.804	0.670	0.613
	Rat	0.648	0.756	0.643	0.646
	Mouse	0.490	0.857	**0.683**	0.570
Single modal (only smiles)	Human	0.515	**0.839**	**0.694**	0.591
	Rat	0.534	**0.825**	**0.667**	0.593
	Mouse	0.436	**0.868**	0.666	0.527
Our model	Human	**0.644**	0.743	0.640	**0.626**
	Rat	**0.720**	0.695	0.607	**0.659**
	Mouse	**0.583**	0.769	0.604	**0.595**

Appendix table 1. Cross-species performance under alternative modeling settings. A set of evaluation metrics was employed to assess the performance of both models, including sensitivity, specificity, precision, and F1 score.