Integrating Bilinear Transduction with Message Passing Neural Networks for Improved ADMET Property Prediction

Claire Suen[1]   **Alan C. Cheng**[1]

Abstract

With the increasing application of machine learning and deep learning in drug discovery comes the significant challenge of addressing censored molecular property datasets. Pharmaceutical assays frequently generate censored data where measurement limitations prevent recording exact values beyond predetermined thresholds. Standard deep learning approaches struggle with this censoring, often producing systematic prediction errors even for in-distribution molecules. Building on the established strengths of Chemprop and the concept of bilinear transduction, we present a method that integrates bilinear transduction into Chemprop's message-passing neural network. This integration allows us to effectively leverage domain-specific structural relationships between molecules, addressing current limitations in molecular property prediction. Our comprehensive evaluation across multiple ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties demonstrate that our method outperforms standard D-MPNN baselines, with improvements exceeding 100% for heavily censored datasets like CYP2C9 inhibition and CYP2D6 inhibition. This practical solution requires no additional experimental measurements while improving ADMET property prediction accuracy, particularly in the challenging high-censoring regimes common in pharmaceutical research.

1. Introduction

Deep learning approaches have revolutionized molecule property prediction in drug discovery (Cáceres et al., 2020;

Hessler & Baringhaus, 2018; Shen & Nicolaou, 2019; Walters & Barzilay, 2021; Ferreira & Andricopulo, 2019), but a significant challenge remains with censored datasets (Figure 6). Drug discovery is often hampered by censored labels, where concentration levels fall outside the assay's limit of detection (Lind, 2010; Svensson et al., 2025; Sheridan et al., 2020). A prime example is CYP3A4 inhibition assays, where approximately 60% of the data is censored due to predefined concentration ranges. This extensive censoring significantly impacts model performance, often resulting in systematic over- or under-prediction of molecules near the threshold value when their true values lie outside the censored region. Such errors dilute predictive capabilities even for in-distribution molecules that should be well-characterized by the model. Importantly, these censored labels contain valuable information that, when properly leveraged, can significantly improve the non-censored region accuracy of predictive models (Svensson et al., 2025).

1.1. Limitations of Deep Learning Models on Censored Data

With graph neural network models like Chemprop (Yang et al., 2019), which employ a directed message passing neural network (D-MPNN) architecture to predict properties from molecular graphs, the models tend to predict values at or near the censoring threshold rather than the true in-distribution value. A possible hypothesis for this behavior is that the model learns to associate certain molecular features with specific threshold values. However, removing censored values is not an effective solution, as it eliminates important examples for training and reduces overall performance (Svensson et al., 2025).

1.2. Related Work

Recent advances in deep learning offer promising approaches to address this challenge. (Segal et al., 2025) demonstrated that bilinear transduction—a method that leverages analogical input-target relationships—can significantly improve out-of-support (OOS) property prediction for both materials and molecules. By reformulating the prediction problem to learn from differences between compounds rather than absolute properties, bilinear transduction poses a
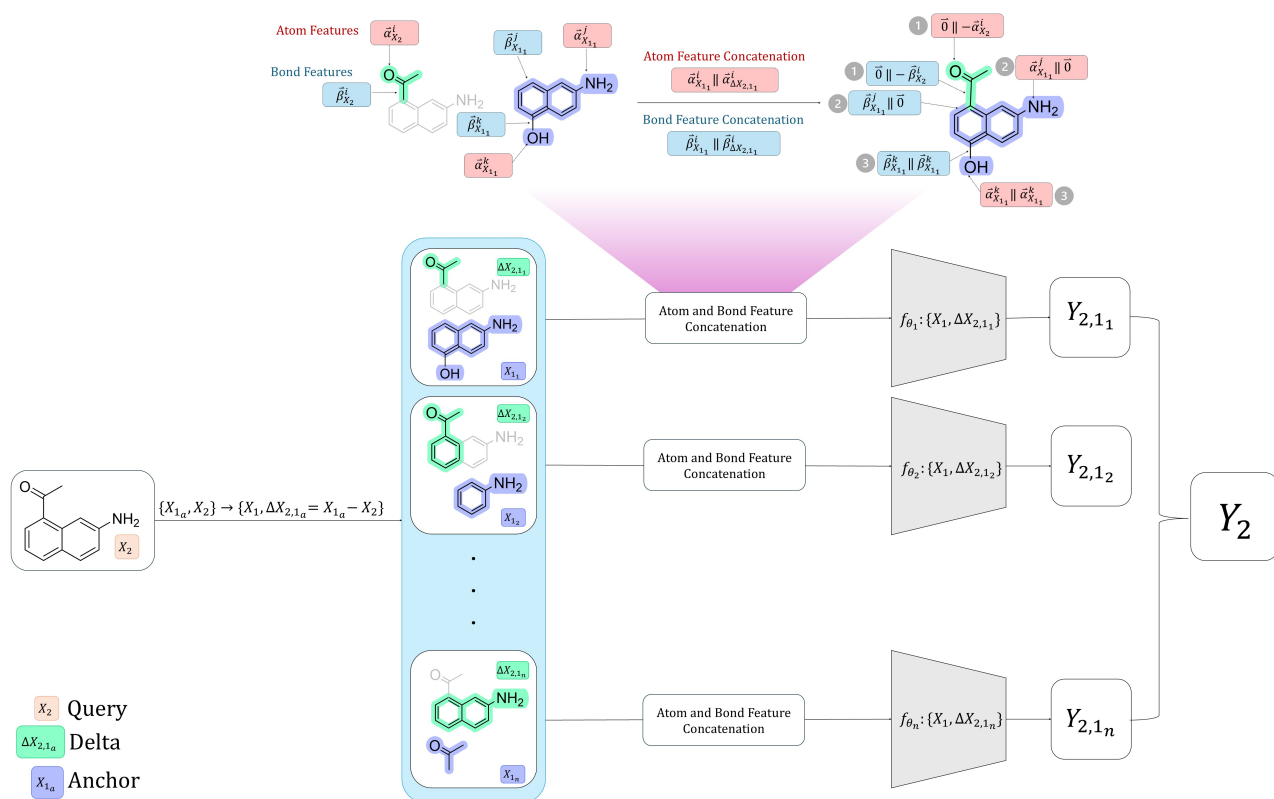
---

[1]Modeling and Informatics, Discovery Chemistry, Merck & Co., Inc., South San Francisco, Ca, USA. Correspondence to: Claire Suen <claire.suen@merck.com>, Alan C. Cheng <alan.cheng@merck.com>.

Figure 1. Bilinear transduction representation and implementation within Chemprop message-passing feature concatenation architecture. The model handles three distinct molecular (atom & bond) transformation scenarios: (1) components present only in the query molecule but absent in the anchor (elimination scenario, where anchor features are null and Δ vector contains negative query features), (2) components shared between molecules (preservation scenario, where anchor features are maintained and Δ vector approaches zero), and (3) components unique to the anchor molecule (formation scenario, where both anchor features and Δ vector contain positive values). N anchors are chosen and N models are trained, enabling effective property prediction through ensemble averaging of transformations.

potential method to address the challenge of in-distribution prediction accuracy for ADMET properties.

Beyond single-molecule property prediction, other research has explored models that take multiple molecular inputs to predict a specific outcome, a strategy akin to our bilinear transduction approach. For instance, Grambow (2020) developed a template-free deep learning model using a Directed Message Passing Neural Network (D-MPNN) to predict activation energies from both reactant and product graphs, leveraging atomic difference fingerprints for reaction encoding. Similarly, Chung (2022) introduced DirectML, a D-MPNN-based model that directly predicts solvation free energy and enthalpy for a given solvent-solute pair. These approaches highlight situations where models utilize explicit relationships or transformations between molecular entities to enhance predictive accuracy, providing context for our method that learns from differences between anchor-query pairs to predict a delta using a D-MPNN framework.

Previous work has been done in fixing the issue of data censorship in machine learning (De Santi et al., 2025; Vock et al., 2016; Dănăilă & Buiu, 2024). For example, the Tobit model has been used as a foundation for Deep Tobit Networks that leverage deep neural networks to capture complex nonlinear patterns in censored data (Zhang et al., 2021). However, these approaches generally still struggle with datasets where more than 50% of values are censored. Additionally, these methods focus on reducing prediction bias, quantifying uncertainty, or on out-of-distribution prediction rather than specifically improving in-distribution accuracy, presenting an opportunity for this approach we detail here.

1.3. Bilinear Transduction Approach

Bilinear transduction offers several advantages for handling censored data in ADMET property prediction (Netanyahu et al., 2023). By facilitating learning between anchor-query

2

pairs, knowledge can be transferred across molecules, helping to compensate for sparsity in specific properties. Specifically, bilinear transduction operates on a foundation of relational reasoning rather than absolute value prediction. The relationships between molecular structures and how they transform property values can effectively circumvent the inaccuracy introduced by the censoring boundaries. This approach can create a more robust representation space where the model can learn the underlying property distribution rather than memorization. This approach may also, in the future, provide insights into which molecular features influence specific ADMET property changes, thus enhancing model interpretability (Xu et al., 2019; Zhao et al., 2022).

Unlike previous bilinear transduction implementations that function as standalone architectures, our novel approach integrates bilinear transduction directly within Chemprop's message-passing framework at the feature concatenation stage for both atoms and bonds, as illustrated in Figure 1. By embedding this structure relationship framework within Chemprop we maintain Chemprop's strong molecular representation learning capabilities while adding the ability to effectively reason about property differences across censored datasets.

1.4. Main Findings

In this work, we propose a novel integration of bilinear transduction within Chemprop. We focus on improving in-distribution prediction accuracy for censored datasets. Our main contributions are summarized as follows:

- We integrate bilinear transduction within the Chemprop framework, creating an approach that leverages analogical input-target relationships for molecule property prediction.

- We demonstrate that our approach significantly improves in-distribution prediction accuracy, particularly for datasets with censored labels where measurement limitations prevent recording exact values.

- We provide comprehensive evidence that while standard Chemprop models can handle small percentages of censored data, they degrade significantly at $\sim 50\%$ censoring, predicting values clustered around the censoring threshold even when true values lie well above.

- We evaluate our approach across multiple pharmaceutical assays (including CYP enzymes, CaV 1.2, and hERG MK499) and synthetic datasets to demonstrate consistent performance improvements with our bilinear transduction (BT) approach.

- We analyze the impact of different anchor selection strategies, showing that performance improvement

plateaus around 8-10 anchor molecules while variance continues to decrease with larger anchor sets.

2. Results

2.1. Performance of Bilinear Transduction

Table 1, 2, and 3 present the comparative performance of BT model against Chemprop (D-MPNN) across various properties, levels of censorship, and dataset sizes. For the internal datasets we employed temporal train/test splits (Cáceres et al., 2020; Sheridan, 2013) and for the public datasets a random split was used. More information about the assays and datasets can be found in Table 4, 5, and 6. Our results demonstrate that BT consistently outperforms D-MPNN across both internal and public datasets.

2.1.1. INTERNAL DATA (NATURALLY CENSORED ASSAYS)

The results shown in Table 1 show consistent improvements in both R^2 and RMSE metrics for CYP 3A4 when comparing the BT method to the D-MPNN method across all three training dataset sizes. The largest absolute improvement occurred in the 100k training dataset size, where R^2 increased from 0.22 to 0.32. CYP 2D6 showed even more substantial gains across all training sizes. The 50k training size showed a notable improvement where R^2 increased from -0.02 to 0.19. CYP 2C9 had significant improvements for the 100k and 225,026 training sizes, with R^2 increasing from 0.01 to 0.19 in the 100k training size. RMSE also showed significant improvements across these properties. CaV 1.2 showed improvement with the increased training sizes for all three metrics. hERG MK499 showed modest gains with comparable R^2 values between the two methods.

2.1.2. INTERNAL DATA (SYNTHETIC CENSORED ASSAYS)

When evaluated on synthetically censored datasets (Table 2), BT showed consistent results with the largest absolute improvements observed at higher censoring levels (50th and 75th percentile). For rat P-gp across the three censoring thresholds (25th, 50th, and 75th percentile), the BT method yielded significantly better R^2 values than the D-MPNN method. At the 50th percentile, R^2 increased from 0.18 with D-MPNN to 0.40 with BT. Rat $F_{u,p}$ showed a similar pattern where improvements were most pronounced at higher censoring thresholds. At the 75th percentile for rat $F_{u,p}$, R^2 increased from 0.27 to 0.43. For both assays, the 25th percentile threshold showed smaller absolute differences between models, with the 50th and 75th percentiles demonstrating increasingly substantial improvements in performance metrics.

Table 1. Performance of Bilinear Transduction Models and Baselines on Naturally Censored Internal Datasets. Results are presented across two metrics: R^2 (higher is better) and RMSE (lower is better), with mean $\pm$ standard deviation shown for each. **Bold values** indicate statistically significant better performance where one model outperforms the other greater than one standard deviation. Underlined values indicate cases where performance is statistically equivalent (overlapping within standard deviation) or where differences are within standard deviation ranges.

ASSAY	TRAINING LABEL	$R^2 \uparrow$		RMSE $\downarrow$	
		BT	D-MPNN	BT	D-MPNN
CYP 3A4	50K	0.30±0.03	0.26±0.06	0.47±0.01	0.48±0.02
	100K	**0.32±0.04**	0.22±0.04	**0.46±0.01**	0.49±0.01
	224,593	0.40±0.04	0.29±0.07	**0.43±0.01**	0.47±0.02
CYP 2D6	50K	**0.19±0.05**	-0.02±0.06	**0.45±0.01**	0.51±0.02
	100K	**0.30±0.04**	0.12±0.04	**0.42±0.01**	0.47±0.01
	221,745	**0.32±0.04**	0.11±0.03	**0.41±0.01**	0.47±0.01
CYP 2C9	50K	**0.08±0.03**	-0.07±0.10	0.50±0.01	0.54±0.03
	100K	**0.19±0.04**	0.01±0.04	**0.47±0.01**	0.52±0.01
	225,026	**0.24±0.06**	0.05±0.06	**0.46±0.02**	0.51±0.02
CAV 1.2	50K	0.10±0.06	0.05±0.06	0.27±0.01	0.28±0.01
	100K	**0.11±0.05**	-0.06±0.05	0.27±0.01	0.29±0.01
HERG MK499	50K	0.18±0.05	0.16±0.03	0.48±0.01	0.48±0.01
	100K	0.23±0.04	0.25±0.04	0.46±0.01	0.46±0.01

Table 2. Performance of Bilinear Transduction Models and Baselines on Synthetically Censored Internal Datasets.

ASSAY	TRAINING LABEL	$R^2 \uparrow$		RMSE $\downarrow$	
		BT	D-MPNN	BT	D-MPNN
P-GP, RAT	25TH	**0.57±0.02**	0.50±0.01	**0.35±0.01**	0.37±0.00
	50TH	**0.40±0.05**	0.18±0.12	**0.29±0.01**	0.34±0.02
	75TH	**-0.03±0.06**	-0.50±0.26	**0.22±0.01**	0.26±0.02
RAT $F_{u,p}$	25TH	0.57±0.02	0.56±0.01	0.30±0.01	0.31±0.00
	50TH	0.51±0.02	0.47±0.02	0.23±0.01	0.24±0.00
	75TH	**0.43±0.02**	0.27±0.04	**0.16±0.00**	0.18±0.00

Table 3. Performance of Bilinear Transduction Models and Baselines on Censored Public Datasets.

ASSAY	TRAINING LABEL	$R^2 \uparrow$		RMSE $\downarrow$	
		BT	D-MPNN	BT	D-MPNN
MS (HUMAN)	BASE	0.24±0.04	0.33±0.06	0.46±0.01	0.43±0.02
	50TH	**0.21±0.05**	0.08±0.07	0.38±0.01	0.41±0.02
	75TH	**0.03±0.04**	-0.06±0.02	0.29±0.01	0.30±0.00
MS (RAT)	BASE	0.48±0.02	0.48±0.02	0.46±0.01	0.46±0.01
	50TH	0.14±0.04	0.09±0.02	0.35±0.01	0.36±0.00
	75TH	**0.06±0.02**	-0.56±0.12	**0.23±0.00**	0.30±0.01
CYP 3A4	4,403	0.52±0.02	0.53±0.02	**0.58±0.01**	0.61±0.01

2.1.3. PUBLIC DATA

The public datasets follow the same trend of improvement (Table 3), though with smaller absolute differences, potentially due to their smaller training dataset sizes. For the microsomal stability (human and rat) datasets, BT generally outperforms D-MPNN. For CYP 3A4, BT was found to be around the same performance as D-MPNN although still following a similar trend as the CYP 3A4 internal results.

2.1.4. PERFORMANCE ADVANTAGES IN HIGHLY CENSORED DATA

All performance metrics for BT were found to be equal to or better than D-MPNN. The most significant performance gains were observed for CYP enzyme prediction tasks, particularly for CYP 2D6 and CYP 2C9, where improvements exceeded 100% in some cases. The performance of BT compared to D-MPNN on synthetically censored datasets indicates that BT has the potential for outperformance in scenarios with extensive censoring. These results collectively validate our hypothesis that leveraging analogical input-target relationships through bilinear transduction improves in-distribution prediction accuracy for censored ADMET datasets.

2.2. Optimal Anchor Size

Our analysis of anchor ensemble sizes revealed that performance metrics (R^2 and RMSE) plateau around eight anchors, suggesting this as an optimal number for practical implementation of BT. Figure 2 shows the performance plot for CYP 3A4 (internal). Other assays display a similar pattern of results. While performance improvements became negligible beyond this threshold, the variance in predictions continued to decrease with larger anchor sets, indicating that additional anchors still contribute to model stability even when they no longer substantially improve accuracy. This suggests there is a balance of computational efficiency and predictive performance for drug discovery workflows.

2.3. Anchor Error Distribution

Analysis of the anchor error distributions showed how prediction errors correlate between anchor molecules (Kuncheva & Whitaker, 2003; Dieterich, 2000). Across the assays the errors showed instances of high R^2 (≈ 0.93) and moderate R^2 (≈ 0.81), as shown in Figure 3. The Tanimoto similarities (TS) were compared between the anchor pairs with high error R^2 values and low error R^2 values (Bajusz et al., 2015). Overall, the TS between anchor pairs showed little correlation with their error distribution similarity. Pairs with highly similar error patterns ($R^2 > 0.92$) had TS ranging from 0.10 to 0.13, indicating that structural similarity does not necessarily predict similar error patterns.
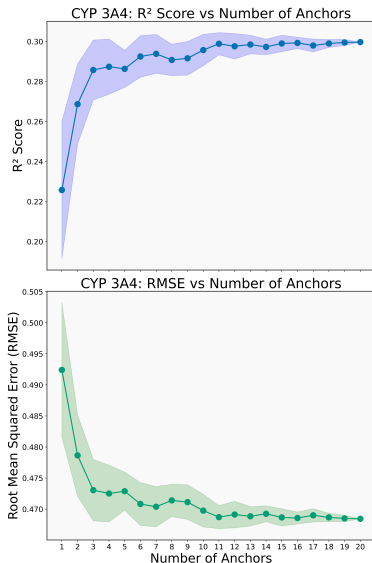

Figure 2. CYP 3A4 model performance metrics across varying anchor ensemble sizes. The plots show individual metrics (top to bottom): R^2 and RMSE values as functions of anchor count. Shaded regions represent standard deviation across sampled anchor combinations at each anchor count (up to 20 combinations sampled).

3. Methods

3.1. Datasets

Ten total ADMET assay datasets were used to compare model performance of Bilinear Transduction and D-MPNN as listed in (Table 4, Table 5, Table 6). Among the ten datasets, seven datasets were collected from internal data from Merck & Co., Inc. (Rahway, NJ, USA). Among the seven datasets, five are naturally censored datasets: CYP 3A4, CYP 2D6, CYP 2C9, CaV 1.2, and hERG MK499. The remaining two datasets are synthetically censored datasets where a synthetic threshold is imposed at three levels (25th percentile, 50th percentile, and 75th percentile) for rat P-gp and rat $F_{u,p}$. Additionally, our methods were tested on three datasets collected from the literature (Kim et al., 2023; Fang et al., 2023): microsomal stability (human), microsomal stability (rat), and CYP 3A4.

3.2. Bilinear Transduction

3.2.1. BILINEAR TRANSDUCTION REPRESENTATION

Bilinear transduction relies on reparametrizing the data into a distribution of the differences. Given a set of molecule structures (X) and ADMET properties (Y), we typically train a D-MPNN model ($f_\theta : X \to Y$). Bilinear transduction relies on the difference's distribution instead. We can reparametrize the problem as such. Let $\Delta X = \{x_1 - x_2 \mid$
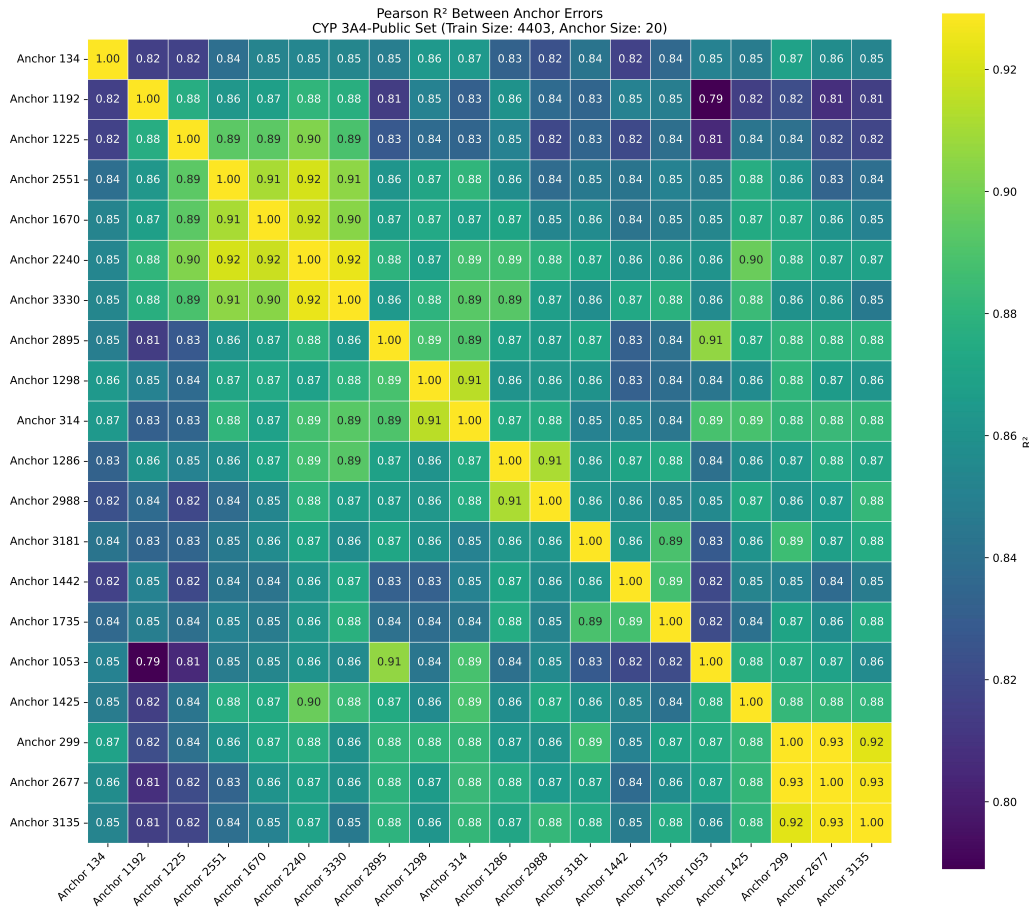
Figure 3. Heatmap of error distribution correlations between anchor molecules for CYP3A4 prediction. The color intensity represents the R^2 coefficient between prediction error distributions for each anchor pair, with darker blue indicating stronger correlation. Clusters of highly correlated anchors ($R^2 > 0.90$) suggest that groups of molecules produce similar error patterns despite structural differences.

$x_1,\ x_2 \in X$} where x_2 is referred to as a query point and x_1 an anchor point.

### 3.2.2. BILINEAR TRANSDUCTION WITHIN CHEMPROP ARCHITECTURE

The bilinear transduction mechanism is integrated into the feature concatenation process for both atoms and bonds. Each molecular component possesses a distinct feature vector representation. For this concatenation procedure, we combine the anchor feature vector with a difference (Δ) vector that encodes the transformation between molecules (Heid & Green, 2022).

This representation handles three distinct scenarios: For components present exclusively in the query (reactant) but absent in the anchor (product), the anchor feature vector is null while the Δ vector contains negative values equal to the query component's features, representing elimination. For

components shared between anchor and query molecules, the representation preserves the anchor features while the Δ vector approaches zero, indicating preservation. For components unique to the anchor molecule, both the anchor feature vector and the Δ vector contain positive values derived from the anchor component, signifying formation of new structural elements.

3.2.3. BILINEAR TRANSDUCTION IMPLEMENTATION

Under the reparameterization discussed in Section 3.2.1 and representation concatenation discussed in Section 3.2.2, we will train $n = 20$ predictors of the form: $f_{\theta_i} : (x_1, \Delta X_{2,1_i}) \rightarrow \Delta Y_{2,1_i}$. To obtain the final prediction ADMET property ($\hat{y}_i$) for a given compound (x_i) we obtain $\hat{y}_i = x_1 - \Delta \hat{x}_{2,1_i}$ for each predictor. Finally, we take $\frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$ to obtain the final prediction. The anchors were chosen randomly from the training set and the same anchors were used for the test set.

Figure 4. **(Left)** Molecular structures of anchor pairs with highly similar error distributions for Public CYP 3A4. This table displays the five anchor pairs exhibiting the highest error correlation coefficients ($R^2 > 0.92$) in CYP3A4 inhibition prediction. For each pair, the molecular structures, Tanimoto similarity coefficient (TS), and R^2 value for error correlation are presented. **(Right)** Molecular structures of anchor pairs with highly dissimilar error distributions for Public CYP 3A4. This table displays the five anchor pairs exhibiting the highest error correlation coefficients ($R^2 < 0.82$) in CYP3A4 inhibition prediction.

3.2.4. BASELINE D-MPNN IMPLEMENTATION

For our comparative analysis, we implemented the directed message passing neural network (D-MPNN) as described in the original Chemprop architecture (Yang et al., 2019). We trained each model using a 2-fold cross-validation approach with 2 ensemble members per fold, resulting in 4 total model predictions per molecule. The predicted value for each test compound was calculated by averaging the 4 individual predictions, and the standard deviation was determined by computing performance metrics independently for each prediction (Kuncheva & Whitaker, 2003; Dietterich, 2000). Hyperparameters for the D-MPNN implementation are detailed in Table 7.

4. Discussion

4.1. Predictive Performance of Bilinear Transduction

Leveraging information from analogical input-target relationships via bilinear transduction (BT) significantly improves the prediction of ADMET property values for in-distribution molecules of censored datasets. While standard

Chemprop models exhibit a degree of robustness to low levels of censoring, their predictive accuracy deteriorates considerably as the percentage of censored data increases. This is exemplified in the synthetically censored datasets where Chemprop's performance degrades significantly once 50% of the dataset is censored (Figure 14, Figure 15). We also notice this pattern in the hERG MK499 assay, which has a relatively low censoring rate (19%), showed similar performance between Chemprop and BT.

Under extreme censored scenarios Chemprop tends to produce predictions clustered around the censoring threshold as seen in the CYP 2C9 (Figure 5), CYP 2D6 (Figure 7), and P-gp rat (50th percentile) parity plots (Figure 8), failing to accurately capture the true range of property values. In contrast, the BT model exhibits the ability to discern the underlying structure-property relationships and remains robust to high levels of censoring. By explicitly modeling these analogical relationships, BT appears to capture a more nuanced understanding of the data compared to standard Chemprop models.
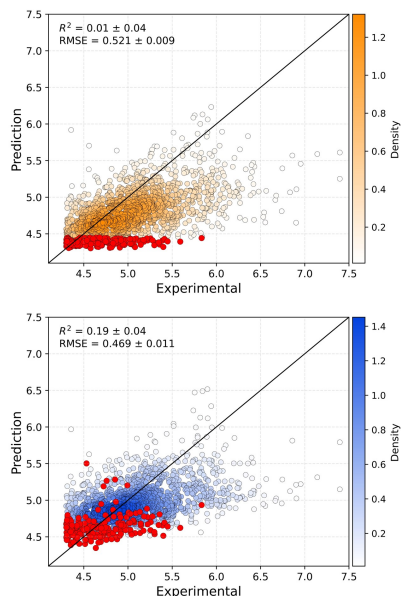
Figure 5. Parity plots for CYP 2C9 inhibition predictions (*n*=100k).(**Top**) D-MPNN baseline. (**Bottom**) BT model with 20 anchors. Red points highlight identical molecules in both plots that D-MPNN clusters near the censoring threshold (4.3003), while BT distributes these same compounds more accurately relative to their true values. This demonstrates BT's superior performance in handling censored data compared to the baseline approach.

4.2. Impact of Anchor Set Characteristics

The size and diversity of anchors chosen can influence the consistency and performance of the BT model. Increasing the number of anchor molecules led to a notable decrease in the variance of predictive performance across different anchor sets decreased. This highlights the importance of a sufficiently large and diverse set of anchors to provide a robust representation of the data's underlying structure and analogical relationships. However, we also observed a plateau in performance gains around 8-10 anchors across the tested assays (Figure 2). This suggests a point of diminishing returns, where adding more anchors does not significantly contribute to further improvements in predictive accuracy. This finding has practical implications for the efficient implementation of BT, indicating that a carefully selected set of around 10 anchors may offer a good balance between computational cost and predictive performance.

4.3. Influence of Anchor Error Correlations

To further understand the influence of the chosen anchor set, the relationship between the error distribution of individual anchors and the predictive performance of the BT model was examined (Figure 3). Our analysis of error cor-

relations between different anchors did not reveal a significant correlation with predictive performance. While some anchors exhibited highly correlated prediction errors, and others showed lower correlations, the overall predictive accuracy achieved using the most and least correlated sets of 5 anchors was comparable. This suggests that the individual error patterns of anchors, while potentially informative about the model's learning process, do not directly dictate the overall predictive power of the BT approach.

Additionally, the average Tanimoto similarity between each pair of anchors of the top 5 most similar error correlations and top 5 most dissimilar error correlations was low, ranging from 0.11 to 0.15, indicating the similar errors generated by two anchors isn't necessarily related to their structural similarity. Instead, it appears that the collective information captured by the set of anchors, regardless of their individual error correlations, appears to be the primary driver of the observed performance improvements.

4.4. Future Directions

We aim to enhance our bilinear transduction approach by developing a multi-anchor learning framework. This framework will leverage multiple anchor molecules during training, moving beyond the current model's separate training for each anchor. This could help improve computational efficiency as well as allow for cross-anchor information gains. Additionally, it would be interesting to extend our approach to non-censored ADMET properties and other modalities within the chemical and biological domains.

4.5. Limitations

Bilinear transduction has shown usefulness in improving censored ADMET property prediction, but there are some areas that can be further explored in future work. For instance, the BT method currently requires additional computational resources due to the need to train a separate model for each anchor molecule, suggesting an opportunity for optimization in future iterations. Additionally, while the method improves predictions for highly censored datasets, there is still an overall challenge in these scenarios where overall predictive power could be further strengthened. Future research could also investigate methods to enhance the interpretability of molecular features that correspond to ADMET property changes, as current analyses show limited correlation between structural similarities and error distribution patterns. Furthermore, exploring ways to maximize performance gains on smaller datasets could also be a valuable direction, as the method's absolute improvements were less pronounced in these cases.

Impact Statement

Improving the prediction of ADMET properties, particularly for heavily censored datasets commonly seen in pharmaceutical research, will empower drug discovery teams to more accurately assess drug candidates, leading to better selection and potentially faster development of effective therapeutics. We have developed an approach integrated within an existing model architecture that significantly enhances the performance of ADMET property prediction in the presence of censoring, offering a practical solution without requiring additional experimental data.

References

Adrian, M., Chung, Y., and Cheng, A. C. Denoising drug discovery data for improved absorption, distribution, metabolism, excretion, and toxicity property prediction. *Journal of Chemical Information and Modeling*, 64(16): 6324–6337, 2024. doi: 10.1021/acs.jcim.4c00639.

Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7:20, 2015.

Cáceres, E. L., Tudor, M., and Cheng, A. C. Deep learning approaches in predicting admet properties. *Future Medicinal Chemistry*, 12(22):1995–1999, 2020. URL https://doi.org/10.4155/fmc-2020-0259.

Chung, Y., Vermeire, F. H., Wu, H., Walker, P. J., Abraham, M. H., and Green, W. H. Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy. *Journal of Chemical Information and Modeling*, 62(3):433–446, 2022. doi: 10.1021/acs.jcim.1c01103. URL https://doi.org/10.1021/acs.jcim.1c01103.

De Santi, L. A., Orlandini, F., Positano, V., Pistoia, L., Sorrentino, F., Messina, G., Roberti, M. G., Missere, M., Schicchi, N., Vallone, A., Santarelli, M. F., Clemente, A., and Meloni, A. Explainable survival analysis of censored clinical data using a neural network approach. *BioMedInformatics*, 5(2):17, 2025. doi: 10.3390/biomedinformatics5020017. URL https://doi.org/10.3390/biomedinformatics5020017.

Dietterich, T. G. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pp. 1–15, Berlin, Heidelberg, 2000. Springer.

Dănăilă, V.-R. and Buiu, C. A deep learning approach to censored regression. *Pattern Analysis and Applications*, 27(1):24, 2024.

Fang, C., Wang, Y., Grater, R., Kapadnis, S., Black, C., Trapa, P., and Sciabola, S. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11):3263–3274, 2023. doi: 10.1021/acs.jcim.3c00160.

Ferreira, L. L. G. and Andricopulo, A. D. Admet modeling approaches in drug discovery. *Drug Discovery Today*, 24 (5):1157–1165, 2019.

Grambow, C. A., Pattanaik, L., and Green, W. H. Deep learning of activation energies. *The Journal of Physical Chemistry Letters*, 11(8):2992–2997, 2020. doi: 10.1021/acs.jpclett.0c00500. URL https://doi.org/10.1021/acs.jpclett.0c00500.

Heid, E. and Green, W. H. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling*, 62(9):2101–2110, 2022. doi: 10.1021/acs.jcim.1c00975.

Hessler, G. and Baringhaus, K.-H. Artificial intelligence in drug design. *Molecules*, 23(10):2520, 2018. doi: 10.3390/molecules23102520.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. Pubchem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. URL https://doi.org/10.1093/nar/gkac956.

Kuncheva, L. I. and Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

Lind, P. Qsar analysis involving assay results which are only known to be greater than, or less than some cut-off limit. *Molecular Informatics*, 29(12):845–852, 2010. doi: 10.1002/minf.201000074.

Netanyahu, A., Gupta, A., Simchowitz, M., Zhang, K., and Agrawal, P. Learning to extrapolate: A transductive approach. *arXiv preprint*, arXiv:2304.14329, 2023. URL https://arxiv.org/abs/2304.14329.

Segal, N., Netanyahu, A., Greenman, K. P., Agrawal, P., and Gomez-Bombarelli, R. Known unknowns: Out-of-distribution property prediction in materials and molecules. *arXiv preprint*, arXiv:2502.05970, 2025. URL https://arxiv.org/abs/2502.05970.

Shen, J. and Nicolaou, C. A. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32-33:29–36, 2019. ISSN 1740-6749. doi: 10.1016/j.ddtec.2020.05.

001. URL https://www.sciencedirect.com/science/article/pii/S1740674920300032.

Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4): 783–790, 2013. doi: 10.1021/ci400084k. URL http://dx.doi.org/10.1021/ci400084k.

Sheridan, R. P., Karnachi, P., Tudor, M., Xu, Y., Liaw, A., Shah, F., Cheng, A. C., Joshi, E., Glick, M., and Alvarez, J. Experimental error, kurtosis, activity cliffs, and methodology: What limits the predictivity of quantitative structure-activity relationship models? *Journal of Chemical Information and Modeling*, 60(4):1969–1982, 2020.

Svensson, E., Friesacher, H. R., Winiwarter, S., Mervin, L., Arany, A., and Engkvist, O. Enhancing uncertainty quantification in drug discovery with censored regression labels. *Artificial Intelligence in the Life Sciences*, 7: 100128, 2025. URL https://doi.org/10.1016/j.ailsci.2025.100128.

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., and O'Connor, P. J. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, 61:119–131, jun 2016. doi: 10.1016/j.jbi.2016.03.009.

Walters, W. P. and Barzilay, R. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of Chemical Research*, 54(2):263–270, 2021. doi: 10.1021/acs.accounts.0c00699. URL https://doi.org/10.1021/acs.accounts.0c00699.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Zhang, J., Li, Z., Song, X., and Ning, H. Deep tobit networks: A novel machine learning approach to microeconometrics. *Neural Networks*, 144:279–296, 2021. URL https://doi.org/10.1016/j.neunet.2021.09.003.

Zhao, L., Jin, W., Akoglu, L., and Shah, N. From stars to subgraphs: Uplifting any gnn with local structure awareness. In *International Conference on Learning Representations (ICLR)*, 2022.

A. Further details on Data

A.1. Data

A.1.1. INTERNAL DATA

Six left-censored assays were collected as listed in Table 4: CYP enzyme inhibition (3A4, 2D6, 2C9, and 2C8), CaV 1.2, and hERG MK499. These assays have a naturally occurring censored region due to measurement limitations. Additionally, we used datasets that were not censored. We impose an artificial threshold at different percentiles (25th, 50th, 75th) on the following: rat P-gp and rat $F_{u,p}$ (Table 5). These datasets allow us to analyze the performance of the BT method on real-world applications and how varying levels of censoring relates to the performance of BT and D-MPNN models. Additionally, all train/test splits were done temporally (Sheridan, 2013).

A.1.2. PUBLIC DATA

Two left-censored assays were obtained from Biogen-Polaris (Fang et al., 2023): microsomal stability (human) and microsomal stability (Rat). Only 30% of both assays were censored and we didn't see much difference between the two method's performances so we also impose artificial thresholds at the 50th and 75th percentile (Table 6). One public CYP 3A4 inhibition assay was obtained (Kim et al., 2023). Additionally, all train/test splits were done randomly.

B. Further details on Methods

B.1. Model Parameters for D-MPNN & BT

All BT and D-MPNN models were built on Chemprop architecture (Yang et al., 2019). Chemprop's default hyperparameters were used besides the hyperparameters listed in Table 7.

B.2. Anchor Selection & Aggregation

B.2.1. TRAINING PHASE ANCHOR SELECTION

During training, we employed a random sampling strategy. A set of 20 molecules was sampled from the in-distribution training dataset to serve as the anchor molecules. For each selected anchor molecule, anchor-query pairs were constructed for the in-distribution training set given to each anchor model.

B.2.2. INFERENCE PHASE

During inference on test molecules, each anchor model was used to predict the property difference (delta value) for each anchor-query (test) molecule pair. Given that the property value of the anchor molecule is a known value (as it was selected from the training set), the predicted property value for the test molecule can be derived by adding the predicted delta value to the known anchor property value.
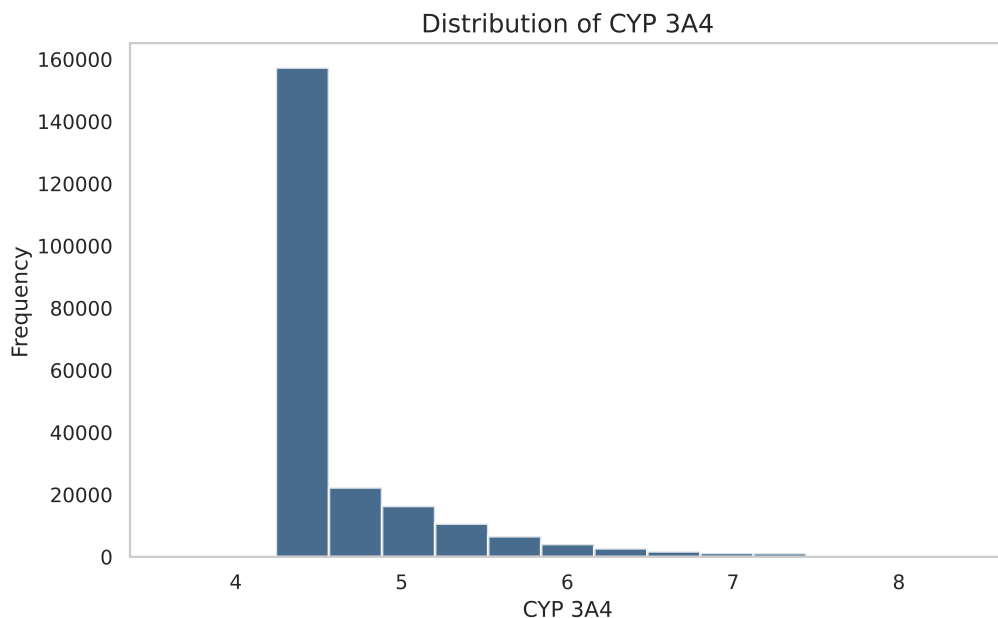
C. Supplemental Figures & Tables

Figure 6. Histogram of the CYP 3A4 Inhibition data. The plot demonstrates the characteristic left-censoring present in the assay, where values lower than the detection limit are not precisely measured.

(a) BT Model

(b) D-MPNN Model
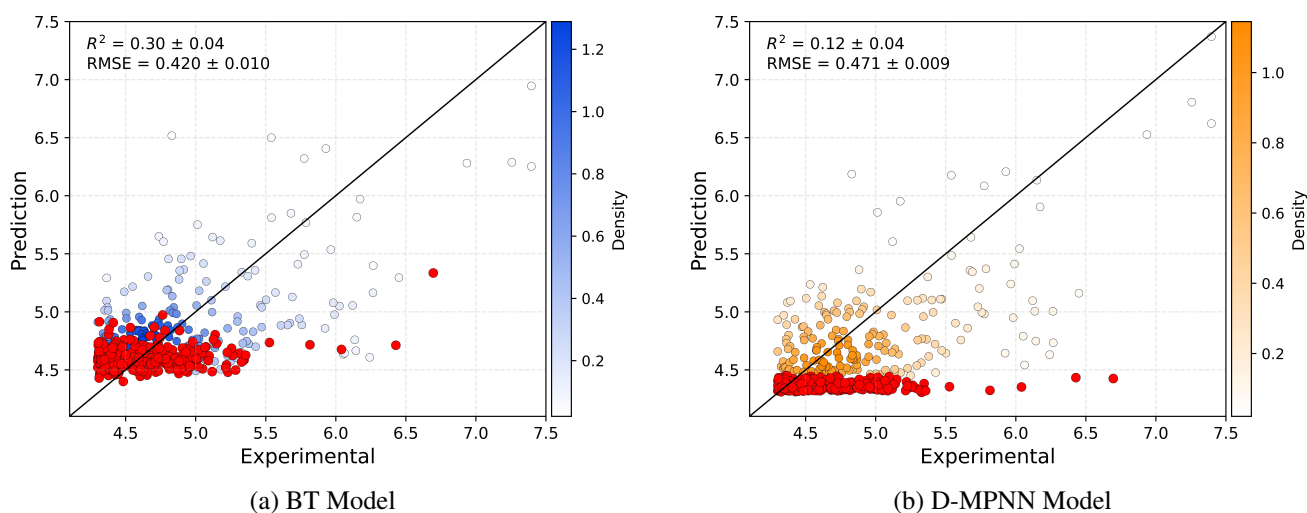
Figure 7. Parity plots for CYP 2D6 inhibition predictions (n=100k). (**Left**) BT model with 20 anchors. (**Right**) D-MPNN baseline. Red points highlight identical molecules in both plots that D-MPNN clusters near the censoring threshold (4.3003), while BT distributes these same compounds more accurately relative to their true values. This demonstrates BT's superior performance in handling censored data compared to the baseline approach.

(a) BT Model

(b) D-MPNN Model

Figure 8. Parity plots for P-gp, rat (50th percentile) predictions (*n*=27,810). (**Left**) BT model with 20 anchors. (**Right**) D-MPNN baseline. Red points highlight identical molecules in both plots that D-MPNN clusters near the censoring threshold (0.65595), while BT distributes these same compounds more accurately relative to their true values. This demonstrates BT's superior performance in handling censored data compared to the baseline approach.
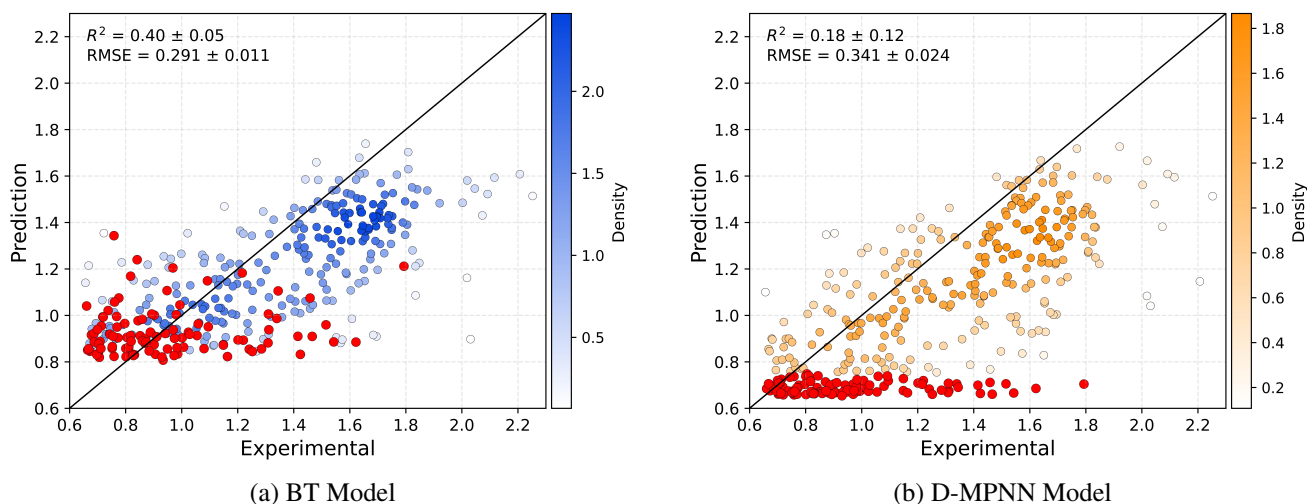

Figure 9. CYP 2D6 model performance metrics across varying anchor ensemble sizes. The plots show individual metrics (top to bottom): R^2 and RMSE values as functions of anchor count. Shaded regions represent standard deviation across sampled anchor combinations at each anchor count (up to 20 combinations sampled).
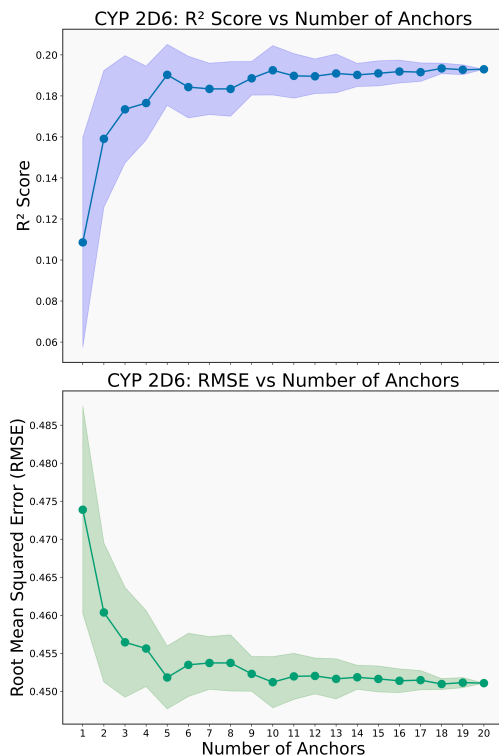
Figure 10. CYP 2C9 model performance metrics across varying anchor ensemble sizes. The plots show individual metrics (top to bottom): R^2 and RMSE values as functions of anchor count. Shaded regions represent standard deviation across sampled anchor combinations at each anchor count (up to 20 combinations sampled).
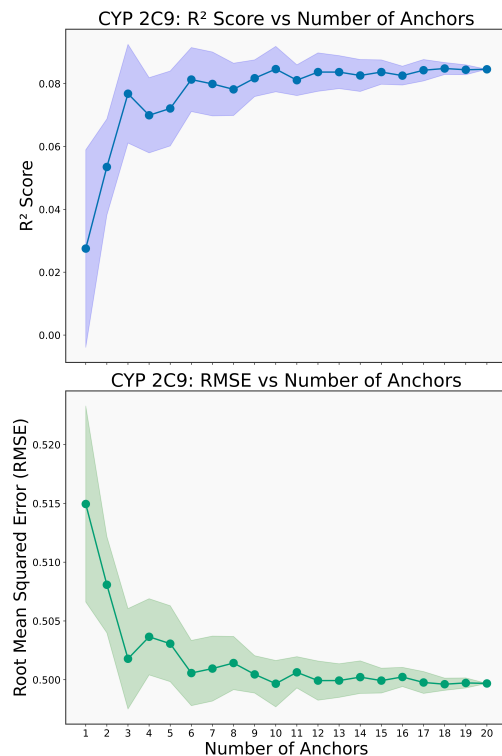
Table 4. Internal (Naturally Censored) Dataset Information.

PROPERTY	TRAINING LABEL	TRAIN SIZE	TEST SIZE	DESCRIPTION	UNITS
CYP3A4	50K 100K 224,593	20,714 41,450 92,940	623	CYP3A4 INHIBITION IC$_{50}$	$-\log_{10}(M)$
CYP2D6	50K 100K 221,745	15,576 31,247 68,918	466	CYP2D6 INHIBITION IC$_{50}$	$-\log_{10}(M)$
CYP2C9	50K 100K 225,026	29,751 59,245 133,132	1,347	CYP2C9 INHIBITION IC$_{50}$	$-\log_{10}(M)$
CAV 1.2	50K 100K	30,850 61,800	652	IC$_{50}$ ON A CA ION CHANNEL	$-\log_{10}(M)$
HERG MK499	50K 100K	40,576 80,826	1,583	HERG CHANNEL IC$_{50}$ THROUGH DISPLACEMENT OF MK-499	$-\log_{10}(M)$
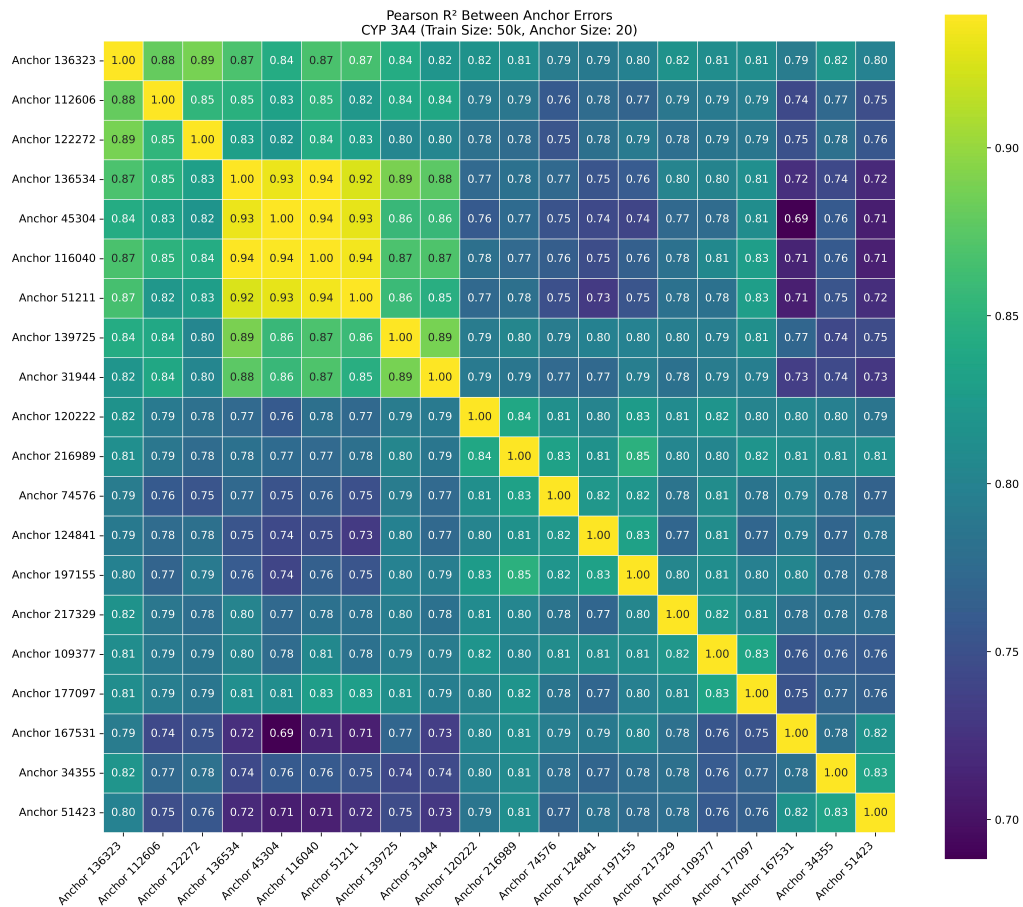
Figure 11. Heatmap of error distribution correlations between anchor molecules for CYP3A4 prediction. The color intensity represents the R^2 coefficient between prediction error distributions for each anchor pair, with darker blue indicating stronger correlation. Clusters of highly correlated anchors ($R^2 > 0.90$) suggest that groups of molecules produce similar error patterns despite structural differences.

Table 5. Internal (Synthetically Censored) Dataset Information.

| PROPERTY	TRAINING LABEL	TRAIN SIZE	TEST SIZE	DESCRIPTION	UNITS
PGP, RAT	25TH 50TH 75TH	20,857 13,905 6,953	577 356 188	RAT BA:AB EFFLUX RATIO	$-\log_{10}(BA:AB\ ratio)$
RAT $F_{u,p}$	25TH 50TH 75TH	40,440 26,964 13,477	2,164 1,620 894	FRACTION OF UNBOUND DRUG IN PLASMA	$-\log_{10}(fraction)$

Figure 12. Heatmap of error distribution correlations between anchor molecules for CYP2D6 prediction. The color intensity represents the R^2 coefficient between prediction error distributions for each anchor pair, with darker blue indicating stronger correlation. Clusters of highly correlated anchors ($R^2 > 0.90$) suggest that groups of molecules produce similar error patterns despite structural differences.

Table 6. Public Dataset Information

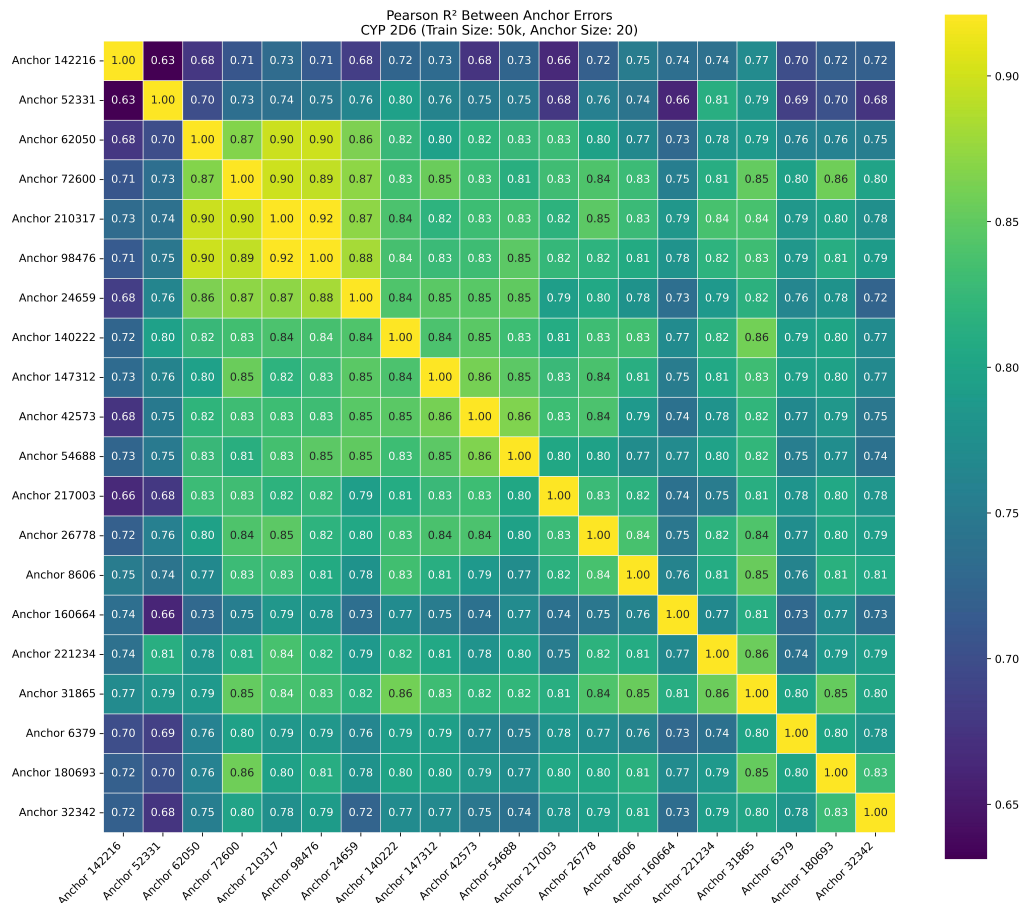| ASSAY	TRAINING LABEL	TRAIN SIZE	TEST SIZE	DESCRIPTION	UNITS	DATA SOURCE
MICROSOMAL STABILITY (HUMAN)	BASE 50TH 75TH	1,715 1,244 633	414 299 139	HUMAN LIVER MICROSOMAL STABILITY REPORTED AS INTRINSIC CLEARANCE	$\log_{10}(\text{ML/MIN/KG})$	FANG ET AL.(2023)
MICROSOMAL STABILITY (RAT)	BASE 50TH 75TH	2,168 1,210 610	540 317 154	RAT MICROSOMAL STABILITY REPORTED AS INTRINSIC CLEARANCE	$\log_{10}(\text{ML/MIN/KG})$	FANG ET AL.(2023)
CYP3A4	4,403	3,344	841	CYP3A4 INHIBITION IC_{50}	$-\log_{10}(nM)$	KIM ET AL.(2023)
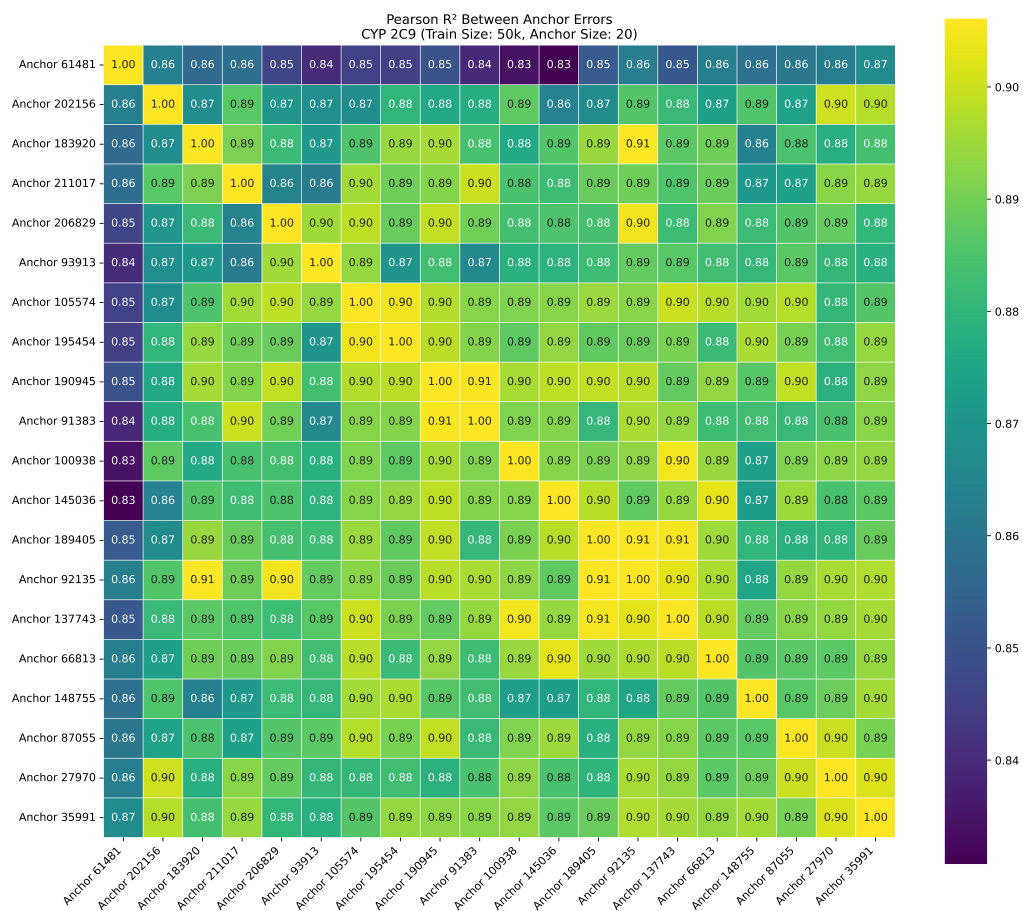
Figure 13. Heatmap of error distribution correlations between anchor molecules for CYP2C9 prediction. The color intensity represents the R^2 coefficient between prediction error distributions for each anchor pair, with darker blue indicating stronger correlation. Clusters of highly correlated anchors ($R^2 > 0.90$) suggest that groups of molecules produce similar error patterns despite structural differences.

Figure 14. Performance comparison of BT and D-MPNN models across censoring levels for internal synthetically censored datasets. The plot shows R^2 scores for Bilinear Transduction (BT, blue) and Directed Message Passing Neural Network (D-MPNN, orange) models on synthetically censored datasets at 25%, 50%, and 75% censoring thresholds. Dashed lines represent P-gp, rat data, while solid lines represent rat $F_{u,p}$ data.

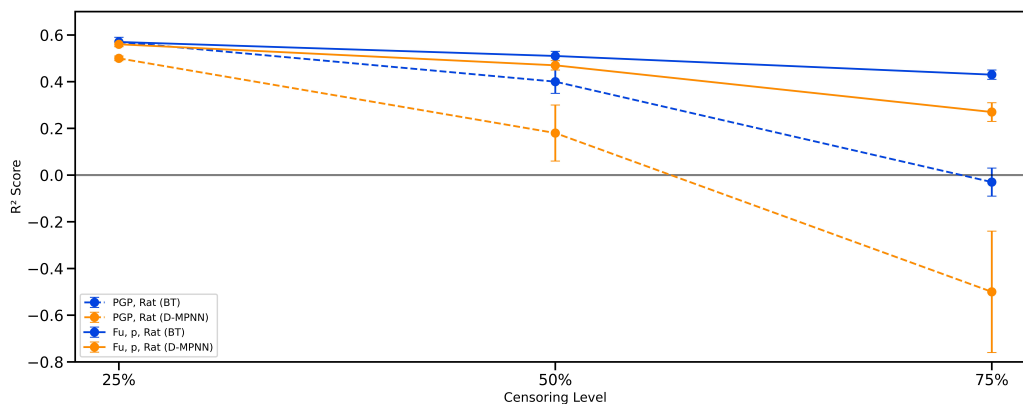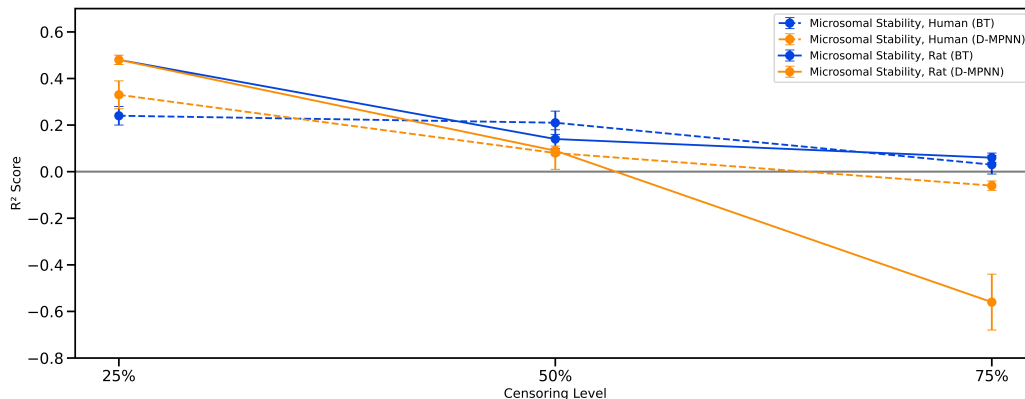*Figure 15.* Performance comparison of BT and D-MPNN models across censoring levels for public synthetically censored datasets. The plot shows R^2 scores for Bilinear Transduction (BT, blue) and Directed Message Passing Neural Network (D-MPNN, orange) models on synthetically censored datasets at 25%, 50%, and 75% censoring thresholds. Dashed lines represent human microsomal stability data, while solid lines represent rat microsomal stability data.

Table 7. Hyperparameter Information. The hyperparameters employed in this work are adopted from (Adrian et al., 2024)

HYPERPARAMETER	VALUE
MPN DEPTH	4
MPN HIDDEN SIZE	600
FFN NUMBER OF LAYERS	4
FFN HIDDEN SIZE	1300
DROPOUT	0
AGGREGATION	NORM
NUMBER OF FOLDS (TRAINING/VALIDATION SPLIT SEED)	2
ENSEMBLE SIZE (PARAMETER INITIALIZATION SEED)	2
EPOCHS	60
NUMBER OF ANCHORS (BT ONLY)	20

Table 8. Performance of Bilinear Transduction Models and Baselines on Naturally Censored Internal Datasets. Results are presented across three metrics: R^2 (higher is better), MAE (lower is better), and RMSE (lower is better), with mean $\pm$ standard deviation shown for each. **Bold values** indicate statistically significant better performance where one model outperforms the other greater than one standard deviation. <u>Underlined values</u> indicate cases where performance is statistically equivalent (overlapping within standard deviation) or where differences are within standard deviation ranges.

ASSAY	TRAINING LABEL	$R^2 \uparrow$		MAE $\downarrow$		RMSE $\downarrow$	
		BT	D-MPNN	BT	D-MPNN	BT	D-MPNN
CYP 3A4	50K	<u>0.30±0.03</u>	<u>0.26±0.06</u>	<u>0.35±0.01</u>	<u>0.35±0.01</u>	<u>0.47±0.01</u>	<u>0.48±0.02</u>
	100K	**0.32±0.04**	0.22±0.04	<u>0.34±0.01</u>	<u>0.35±0.01</u>	**0.46±0.01**	0.49±0.01
	224,593	<u>0.40±0.04</u>	0.29±0.07	<u>0.33±0.01</u>	<u>0.34±0.01</u>	**0.43±0.01**	0.47±0.02
CYP 2D6	50K	**0.19±0.05**	-0.02±0.06	**0.32±0.01**	0.35±0.01	**0.45±0.01**	0.51±0.02
	100K	**0.30±0.04**	0.12±0.04	**0.31±0.01**	0.34±0.00	**0.42±0.01**	0.47±0.01
	221,745	**0.32±0.04**	0.11±0.03	**0.30±0.01**	0.33±0.01	**0.41±0.01**	0.47±0.01
CYP 2C9	50K	**0.08±0.03**	-0.07±0.10	**0.37±0.01**	0.40±0.02	**0.50±0.01**	0.54±0.03
	100K	**0.19±0.04**	0.01±0.04	**0.35±0.01**	0.39±0.01	**0.47±0.01**	0.52±0.01
	225,026	**0.24±0.06**	0.05±0.06	**0.34±0.02**	0.38±0.01	**0.46±0.02**	0.51±0.02
CAV 1.2	50K	<u>0.10±0.06</u>	<u>0.05±0.06</u>	<u>0.21±0.01</u>	<u>0.21±0.01</u>	<u>0.27±0.01</u>	<u>0.28±0.01</u>
	100K	**0.11±0.05**	-0.06±0.05	**0.20±0.00**	0.22±0.01	<u>0.27±0.01</u>	<u>0.29±0.01</u>
HERG MK499	50K	<u>0.18±0.05</u>	<u>0.16±0.03</u>	<u>0.37±0.01</u>	<u>0.36±0.01</u>	<u>0.48±0.01</u>	<u>0.48±0.01</u>
	100K	<u>0.23±0.04</u>	<u>0.25±0.04</u>	<u>0.36±0.01</u>	<u>0.34±0.01</u>	<u>0.46±0.01</u>	<u>0.46±0.01</u>

Table 9. Performance of Bilinear Transduction Models and Baselines on Synthetically Censored Internal Datasets.

ASSAY	TRAINING LABEL	$R^2 \uparrow$		MAE $\downarrow$		RMSE $\downarrow$	
		BT	D-MPNN	BT	D-MPNN	BT	D-MPNN
P-GP, RAT	25th	**0.57±0.02**	0.50±0.01	**0.26±0.01**	0.28±0.00	**0.35±0.01**	0.37±0.00
	50th	**0.40±0.05**	0.18±0.12	**0.23±0.01**	0.27±0.02	**0.29±0.01**	0.34±0.02
	75th	**-0.03±0.06**	-0.50±0.26	**0.17±0.01**	0.21±0.02	**0.22±0.01**	0.26±0.02
RAT $F_{u,p}$	25th	<u>0.57±0.02</u>	<u>0.56±0.01</u>	<u>0.22±0.01</u>	<u>0.22±0.00</u>	<u>0.30±0.01</u>	<u>0.31±0.00</u>
	50th	<u>0.51±0.02</u>	<u>0.47±0.02</u>	<u>0.18±0.00</u>	<u>0.18±0.00</u>	<u>0.23±0.01</u>	<u>0.24±0.00</u>
	75th	**0.43±0.02**	0.27±0.04	**0.12±0.00**	0.13±0.00	**0.16±0.00**	0.18±0.00

Table 10. Performance of Bilinear Transduction Models and Baselines on Censored Public Datasets.

ASSAY	TRAINING LABEL	$R^2 \uparrow$		MAE $\downarrow$		RMSE $\downarrow$	
		BT	D-MPNN	BT	D-MPNN	BT	D-MPNN
MS (HUMAN)	BASE	<u>0.24±0.04</u>	<u>0.33±0.06</u>	<u>0.36±0.01</u>	<u>0.33±0.02</u>	<u>0.46±0.01</u>	<u>0.43±0.02</u>
	50th	**0.21±0.05**	0.08±0.07	<u>0.31±0.01</u>	<u>0.32±0.01</u>	<u>0.38±0.01</u>	<u>0.41±0.02</u>
	75th	**0.03±0.04**	-0.06±0.02	<u>0.24±0.00</u>	**0.23±0.00**	<u>0.29±0.01</u>	<u>0.30±0.00</u>
MS (RAT)	BASE	<u>0.48±0.02</u>	<u>0.48±0.02</u>	<u>0.36±0.01</u>	<u>0.36±0.01</u>	<u>0.46±0.01</u>	<u>0.46±0.01</u>
	50th	<u>0.14±0.04</u>	<u>0.09±0.02</u>	0.30±0.01	**0.28±0.00**	<u>0.35±0.01</u>	<u>0.36±0.00</u>
	75th	**0.06±0.02**	-0.56±0.12	**0.19±0.00**	0.24±0.01	**0.23±0.00**	0.30±0.01
CYP 3A4	4,403	<u>0.52±0.02</u>	<u>0.53±0.02</u>	<u>0.43±0.01</u>	<u>0.44±0.01</u>	**0.58±0.01**	0.61±0.01