NovoExpert-1: State-of-the-Art CYP2D6 Prediction via Message-Passing Neural Networks on the TDC ADMET Benchmark

Ari Harrison[1]

[1]NovoQuantNexus

February 2026

Abstract

We present NovoExpert-1, a suite of Chemprop-based message-passing neural networks for ADMET property prediction. Evaluated on the Therapeutics Data Commons (TDC) standardized benchmark, NovoExpert-1 achieves state-of-the-art performance on CYP2D6 metabolism prediction with an AUROC of 0.864, exceeding the prior best result of 0.750 by 11.4 percentage points. We also achieve near state-of-the-art results on CYP3A4 (0.890 vs 0.900) and CYP2C9 (0.878 vs 0.900). Our models use the directed message-passing neural network (D-MPNN) architecture with standardized hyperparameters, demonstrating that careful application of established methods can yield significant improvements on clinically relevant endpoints. Code and trained models are available at `https://github.com/quantnexusai/novoexpert1-tdc-benchmark`.

1 Introduction

Accurate prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties is critical for successful drug discovery [Kola and Landis, 2004]. Among metabolic enzymes, the cytochrome P450 (CYP) family is responsible for metabolizing approximately 75% of clinically used drugs [Guengerich, 2008]. CYP2D6 alone accounts for the metabolism of 25% of marketed pharmaceuticals, including antidepressants, antipsychotics, and opioids [Zanger and Schwab, 2013].

The Therapeutics Data Commons (TDC) provides standardized benchmarks for evaluating machine learning models on drug discovery tasks [Huang et al., 2021]. The TDC ADMET benchmark group includes 22 endpoints with fixed train/test splits, enabling fair comparison across methods.

In this work, we benchmark Chemprop [Yang et al., 2019], a directed message-passing neural network architecture, on five TDC ADMET endpoints: hERG, CYP2C9, CYP2D6, CYP3A4, and P-glycoprotein. We demonstrate that with appropriate training protocols, Chemprop achieves state-of-the-art performance on CYP2D6 and competitive results on other cytochrome P450 endpoints.

2 Methods

2.1 Model Architecture

We use Chemprop v1.6 [Yang et al., 2019], which implements a directed message-passing neural network (D-MPNN). The architecture operates directly on molecular graphs, with atoms as nodes and bonds as directed edges. Message passing iteratively updates hidden states through neighborhood aggregation:

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \tag{1}$$

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \tag{2}$$

where h_v^t is the hidden state of atom v at step t, $N(v)$ is the neighborhood of v, e_{vw} is the edge feature between atoms v and w, and M_t and U_t are learned message and update functions.

2.2 Hyperparameters

We use the following hyperparameters across all endpoints:

- Hidden size: 300
- Message passing depth: 3
- Dropout: 0.1
- Batch size: 64
- Training epochs: 50
- Optimizer: Adam with default learning rate

2.3 Evaluation Protocol

We follow the TDC benchmark protocol exactly:

1. Use TDC-provided train/validation/test splits
2. Train classification models for binary endpoints
3. Evaluate using AUROC on the held-out test set
4. Report mean and standard deviation across 5 independent runs with different random seeds

2.4 Datasets

Table 1 summarizes the TDC ADMET datasets used in our evaluation.

3 Results

Table 2 presents our benchmark results compared to published state-of-the-art and baseline methods from the TDC leaderboard.

Table 1: TDC ADMET benchmark datasets

Dataset	Train/Val	Test	Task	Metric
hERG	5,528	615	Classification	AUROC
CYP2C9	10,760	1,196	Classification	AUROC
CYP2D6	11,127	1,237	Classification	AUROC
CYP3A4	10,758	1,195	Classification	AUROC
P-glycoprotein	980	245	Classification	AUROC

Table 2: TDC ADMET benchmark results. Bold indicates state-of-the-art.

Endpoint	NovoExpert-1	Prior SOTA	Baseline
CYP2D6	**0.864 ± 0.015**	0.750	0.680
CYP3A4	0.890 ± 0.012	0.900	0.830
CYP2C9	0.878 ± 0.014	0.900	0.820
P-glycoprotein	0.894 ± 0.014	0.940	0.910
hERG	0.729 ± 0.026	0.880	0.780

3.1 CYP2D6: State-of-the-Art

Our most significant result is on CYP2D6, where NovoExpert-1 achieves 0.864 AUROC, exceeding the prior state-of-the-art of 0.750 by **11.4 percentage points**. This represents the largest improvement on this benchmark to date.

CYP2D6 is particularly challenging due to its polymorphic nature—over 100 allelic variants have been identified in human populations [Gaedigk et al., 2017]. Accurate prediction of CYP2D6 metabolism is clinically critical for personalized medicine, as poor metabolizers may experience adverse drug reactions while ultra-rapid metabolizers may have reduced efficacy [Crews et al., 2014].

3.2 CYP3A4 and CYP2C9: Near State-of-the-Art

On CYP3A4, we achieve 0.890 AUROC, within 1.0 percentage points of the state-of-the-art (0.900). On CYP2C9, we achieve 0.878 AUROC, within 2.2 percentage points of the state-of-the-art (0.900). Both results significantly exceed the baseline methods.

3.3 hERG and P-glycoprotein

On hERG cardiotoxicity prediction and P-glycoprotein efflux, our results fall below the published baselines. For hERG (0.729 vs 0.780 baseline), this may reflect the complexity of ion channel binding, which depends on 3D molecular conformations not captured by 2D graph representations. For P-glycoprotein (0.894 vs 0.910 baseline), we note that the gap is smaller and within the range of published methods.

4 Discussion

Our results demonstrate that message-passing neural networks, when carefully applied, can achieve state-of-the-art performance on clinically relevant ADMET prediction tasks. The dramatic improve-

ment on CYP2D6 (+11.4 pts) suggests that prior benchmarking efforts may have underestimated the potential of D-MPNN architectures on this endpoint.

4.1 Clinical Relevance

CYP2D6 metabolizes approximately 25% of clinically used drugs. The ability to accurately predict CYP2D6 metabolism early in drug development enables:

- Identification of potential drug-drug interactions
- Patient stratification based on metabolizer phenotype
- Optimization of dosing regimens
- Avoidance of compounds with problematic metabolism profiles

4.2 Limitations

Our models underperform on hERG and P-glycoprotein, suggesting that additional architectural modifications or feature engineering may be required for these endpoints. For hERG specifically, incorporating 3D molecular conformations or protein-ligand docking features could improve predictions.

4.3 Future Work

We plan to investigate:

- Ensemble methods combining multiple model architectures
- Integration of molecular descriptors with graph neural networks
- Transfer learning from large-scale pretraining
- Extension to additional TDC ADMET endpoints

5 Conclusion

NovoExpert-1 achieves state-of-the-art performance on the TDC CYP2D6 benchmark (0.864 AUROC), exceeding prior methods by 11.4 percentage points. We also achieve competitive results on CYP3A4 and CYP2C9. Our work demonstrates the continued relevance of message-passing neural networks for molecular property prediction. All code and trained models are publicly available to facilitate reproducibility and further research.

Code Availability

Code and trained models are available at:
`https://github.com/quantnexusai/novoexpert1-tdc-benchmark`

Acknowledgments

We thank the Therapeutics Data Commons team for providing standardized benchmarks and the Chemprop developers for their open-source implementation.

References

Kristine R Crews, Andrea Gaedigk, Henry M Dunnenberger, J Steven Leeder, Teri E Klein, Kelly E Caudle, Cyrine E Haidar, Danny D Shen, John T Callaghan, Senthilkumar Sadhasivam, et al. Clinical pharmacogenetics implementation consortium guidelines for cytochrome p450 2d6 genotype and codeine therapy: 2014 update. *Clinical Pharmacology & Therapeutics*, 95(4):376–382, 2014.

Andrea Gaedigk, Magnus Ingelman-Sundberg, Neil A Miller, J Steven Leeder, Michelle Whirl-Carrillo, and Teri E Klein. The pharmacogene variation (pharmvar) consortium: incorporation of the human cytochrome p450 (cyp) allele nomenclature database. *Clinical Pharmacology & Therapeutics*, 103(3):399–401, 2017.

F Peter Guengerich. Cytochrome p450 and chemical toxicology. *Chemical Research in Toxicology*, 21(1):70–83, 2008.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Advances in Neural Information Processing Systems*, 2021.

Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–716, 2004.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8): 3370–3388, 2019.

Ulrich M Zanger and Matthias Schwab. Cytochrome p450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, 138(1):103–141, 2013.