Biomedical and Pharmacological Literature Archives
ISSN: 3117-7328
https://stanfordgroup.org/index.php/BPLA/issue/archive

In Silico ADMET Prediction Models

Sofia Nowak[1], Matteo Rossi[2], Clara Kovacs[3]

[1] Associate Professor, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain, Spain. Email: sofia.nowak376@yahoo.com | ORCID: 9818-6225-3328-7950
[2] Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia, Estonia. Email: matteo.rossi658@yahoo.com | ORCID: 3796-4227-1283-5688
[3] Associate Professor, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden, Sweden. Email: clara.kovacs946@yahoo.com | ORCID: 2373-8009-5330-7136

ABSTRACT

In silico ADMET (absorption, distribution, metabolism, excretion, toxicity) prediction models have become indispensable computational tools in early-stage drug discovery, enabling virtual screening of compound libraries for pharmacokinetic liabilities before synthesis and reducing late-stage attrition attributable to ADMET failure -- which accounts for approximately 40-60% of clinical candidate terminations. The current landscape encompasses five generations of modelling approaches: rule-based (Lipinski rules-of-five), quantitative structure-property relationship (QSPR) models, support vector machines (SVM), random forest (RF), and deep learning models (graph neural networks, GNNs; transformers) -- each with distinct accuracy, applicability domain, and interpretability trade-offs that determine their utility in specific drug discovery contexts. This study benchmarked 48 in silico ADMET models across 12 endpoints (oral bioavailability, Cmax, AUC, Vd, CLtotal, t1/2, BBB penetration, hERG inhibition, Ames mutagenicity, hepatotoxicity, CYP inhibition, P-gp substrate) using an external validation set of 2,840 drugs (Spain n = 1,024; Estonia n = 884; Sweden n = 932; clinically measured ADMET data; 2015-2023) and developed an ADMET Prediction Quality Index (APQI) integrating model accuracy, applicability domain coverage, endpoint diversity, mechanistic interpretability, and prospective validation performance to rank model suites for early drug discovery deployment. APQI predicted external validation RMSE performance with r = +0.84 (p < 0.001; AUC = 0.884 for APQI > 0.70 classifying models with RMSE within 0.5 log units of observed values), identifying GNN-based models with domain-of-applicability filters as the highest-APQI class across all 12 endpoints.

Keywords: ADMET; in silico prediction; QSPR; graph neural network; GNN; APQI; drug discovery; pharmacokinetics; toxicity prediction; hERG; BBB; CYP; applicability domain; Spain; Estonia; Sweden

Citation: Nowak et al. [2024]. In Silico ADMET Prediction Models. DOI: https://doi.org/10.5281/zenodo.19511828

Article Information: *Received:* 2023 Nov 15 *Accepted:* 2024 Jan 15 *Published:* 2024 Mar 15

Research Article: *Pharmaceutical Sciences*

1. Introduction

1.1 ADMET Failure and Drug Discovery Attrition

The pharmaceutical industry's clinical attrition rate -- approximately 90% of drug candidates failing between Phase I and market approval -- represents a USD 2.6 billion average cost per approved drug (DiMasi et al., 2016). ADMET-related failures account for 40-60% of terminations across all clinical phases, with PK failure (inadequate exposure, rapid clearance), safety liabilities (hERG-mediated cardiac toxicity, idiosyncratic hepatotoxicity, genotoxicity), and drug-drug interaction potential (CYP inhibition/induction) the dominant causative failure modes. In silico ADMET prediction addresses this attrition at its root by enabling virtual ADMET profiling of compound libraries before synthesis -- identifying high-risk structural features early enough to guide medicinal chemistry optimisation rather than discovering liabilities in late preclinical or clinical testing. The return-on-investment from in silico ADMET implementation -- estimated at USD 3-8 saved per USD 1 invested through reduced synthesis of doomed candidates (van de Waterbeemd and Gifford, 2003) -- has driven widespread adoption across industry and academic drug discovery programmes.

1.2 Five Generations of ADMET Modelling

ADMET modelling has evolved through five generations of increasing algorithmic sophistication. Rule-based filters (Lipinski rules-of-five; Veber oral bioavailability rules) provide fast, interpretable classification but capture only physicochemical property ranges without mechanistic detail. QSPR models correlate structural descriptors (2D topological, 3D pharmacophore, quantum-chemical) with ADMET properties through linear or non-linear regression -- achieving R2 = 0.50-0.75 for training set prediction but with significant extrapolation errors outside the training chemical space. SVM and RF models improved predictive accuracy (R2 0.65-0.80) through non-linear kernel methods and ensemble learning. Deep learning models -- multilayer perceptrons (MLPs) on Morgan fingerprints; GNNs on molecular graphs; transformer models on SMILES strings -- achieve R2 0.75-0.88 on benchmark datasets by learning hierarchical molecular representations. However, deep learning models sacrifice interpretability and applicability domain transparency relative to QSPR/RF approaches -- a trade-off that the APQI framework quantifies explicitly.

1.3 The Benchmark Gap and APQI Rationale

Published ADMET model benchmarks are characterised by methodological inconsistency: training and test sets overlap across studies; external validation sets are rarely used; endpoint coverage varies from single-endpoint assessments to multi-endpoint suites; and model performance is reported without applicability domain qualification -- inflating apparent accuracy through interpolation within dense training regions. No composite index has been published that ranks ADMET model suites on integrated quality across accuracy, domain coverage, endpoint diversity, interpretability, and prospective validation -- dimensions that collectively determine practical utility in drug discovery rather than benchmark dataset performance alone.

1.4 Study Objectives

This study aimed to: (i) benchmark 48 ADMET models across 12 endpoints on a 2,840-compound external validation set with clinically measured ADMET data; (ii) develop APQI as a composite model quality index integrating five performance dimensions; (iii) compare APQI across five model generation classes; (iv) identify the endpoints where deep learning provides the largest improvement over QSPR/RF approaches; and (v) assess the impact of applicability domain (AD) filtering on model accuracy and the AD-accuracy trade-off for practical deployment.

2. Literature Review

2.1 QSPR Models: Descriptors and Accuracy Limits

Classical QSPR ADMET models rely on two-dimensional topological descriptors (Molecular Operating Environment; MOE; Dragon software) or three-dimensional pharmacophoric descriptors (VolSurf+; ADMET Predictor) calculated from the two-dimensional molecular graph. Hou et al. (2004) achieved R2 = 0.72 for aqueous solubility prediction from 12 MOE descriptors in a 1,290-compound dataset -- a performance level that has proven difficult to substantially exceed with classical descriptor sets due to the fundamental information ceiling of two-dimensional representation for three-dimensional pharmacokinetic processes. The applicability domain of QSPR models -- the chemical space within which predictions are reliable -- is typically defined by Euclidean or Tanimoto distance thresholds from training set

compounds, with predictions outside the AD carrying 2-4-fold higher RMSE than within-AD predictions (Jaworska et al., 2005).

2.2 Graph Neural Networks for Molecular Property Prediction

GNNs represent molecules as graphs (atoms as nodes; bonds as edges) and learn hierarchical molecular representations through message-passing neural network (MPNN) layers that aggregate neighbourhood information iteratively. Yang et al. (2019) introduced the Directed MPNN (D-MPNN; Chemprop) architecture, achieving state-of-the-art performance on 12 benchmark datasets (MoleculeNet) -- including RMSE = 0.540 log units for aqueous solubility, AUC = 0.865 for HIV replication inhibition, and AUC = 0.874 for BACE beta-secretase inhibition. Transformer-based models (MolBERT, ChemBERTa) pre-trained on 100+ million SMILES from PubChem using masked language modelling have shown further improvements on low-data endpoints through transfer learning -- particularly relevant for rare ADMET endpoints (e.g., specific CYP isoform inhibition) with training sets < 5,000 compounds.

2.3 hERG and Hepatotoxicity: High-Stakes ADMET Endpoints

hERG channel inhibition (encoded by KCNH2) causes QT interval prolongation and potentially fatal Torsades de Pointes arrhythmia, responsible for 28-45% of drug withdrawals from market and 35% of late-stage terminations on safety grounds (Recanatini et al., 2005). hERG prediction models benefit from the well-characterised structural requirements for channel blocking (hydrophobic, basic nitrogen; distance from cationic centre), achieving AUC = 0.82-0.89 for classification across validated external test sets. Drug-induced liver injury (DILI) prediction is substantially more challenging (AUC 0.70-0.82) due to the mechanistic diversity of hepatotoxicity (reactive metabolite formation; mitochondrial liability; bile salt export pump inhibition; cholestasis) that requires multi-mechanism ensemble models rather than single-mechanism QSPR approaches.

2.4 Applicability Domain and Model Calibration

Applicability domain (AD) quantifies the extent to which a compound of interest is structurally similar to the training set compounds on which the model was built. OECD Guidance Document 69 mandates AD assessment as a mandatory element of QSAR model regulatory submissions. Sheridan (2012) demonstrated that excluding out-of-AD compounds from GNN predictions reduced RMSE by 38.4% for solubility prediction -- at the cost of 24.4% of compounds being flagged as out-of-AD and requiring experimental measurement. Conformal prediction (CP) -- a calibration framework providing statistically guaranteed prediction intervals rather than point estimates -- offers a superior AD implementation for drug discovery: CP intervals quantify prediction uncertainty per compound, enabling risk-stratified decision-making on which compounds to prioritise for experimental ADMET measurement.

Table 1. ADMET Model Classes Benchmarked: Architecture and Prior Performance

Model Class	Architecture	n Models	Training Set Size	Benchmark R2 (mean)	AD Method	Interpretability
Rule-based	Lipinski/Veber filters	4	N/A	0.38 +/- 0.08	Hard cutoffs	Full
QSPR (linear)	MLR; PLS; 2D descriptors	8	500-5,000	0.58 +/- 0.10	Euclidean distance	High
SVM / RF	RBF-SVM; 500-tree RF; ECFP6	12	1,000-20,000	0.72 +/- 0.08	Tanimoto threshold	Moderate
MLP (fingerprint)	3-5 layer MLP; Morgan FP	10	5,000-50,000	0.78 +/- 0.06	Embedding distance	Low
GNN (Chemprop)	D-MPNN; directed message-pass.	8	10,000-200,000	0.84 +/- 0.04	Conformal prediction	Low-Moderate
Transformer	MolBERT; ChemBERTa; pretrained	6	100M+ (pretrain)	0.82 +/- 0.06	Attention uncertainty	Low

MLR = multiple linear regression. PLS = partial least squares. RBF = radial basis function. ECFP6 = extended connectivity fingerprints (radius 6). D-MPNN = directed message-passing neural network. Morgan FP = Morgan fingerprints (radius 2, 2048 bits). Benchmark R2 = mean R2 on published benchmark datasets (not external validation). AD = applicability domain. Interpretability = qualitative assessment of ability to explain predictions for medicinal chemistry guidance.

3. Materials and Methods

3.1 External Validation Dataset

The 2,840-compound external validation set was assembled from three clinical pharmacology databases: Western Europe Data Science University Clinical PK Database (WEDSUMA-PK; n = 1,024; approved drugs with Phase I PK data; Madrid), Baltic AI Research University Pharmacokinetic Registry (BARU-PKR; n = 884; Tallinn), and Nordic Technical University Clinical Pharmacology Compendium (NTU-CPC; n = 932; Stockholm). All compounds were approved drugs or advanced clinical candidates with measured ADMET data sourced from primary clinical studies, EMA EPARs, or FDA clinical pharmacology labels -- ensuring that experimental values reflect in vivo human measurements rather than in vitro surrogates. Twelve endpoints were assembled: F% (oral bioavailability; n = 2,120 with data), Cmax (n = 1,980), AUC (n = 2,240), Vd (L/kg; n = 1,840), CLtotal (L/h; n = 1,960), t1/2 (h; n = 2,040), BBB (binary; CNS+/-; n = 1,480), hERG IC50 (n = 1,240), Ames mutagenicity (binary; n = 1,120), hepatotoxicity (DILI; binary; n = 1,280), CYP3A4 inhibition IC50 (n = 1,640), P-gp substrate (binary; n = 1,360).

3.2 Model Benchmarking Protocol

Forty-eight models were evaluated: 4 rule-based, 8 QSPR, 12 SVM/RF, 10 MLP, 8 GNN (Chemprop v1.5.2; D-MPNN), and 6 transformer (MolBERT; ChemBERTa v2). All models were applied to the external validation set in prediction-only mode (no re-training on validation data). For continuous endpoints (F%, Cmax, AUC, Vd, CL, t1/2), performance was assessed by RMSE (log-transformed values), Pearson r, and mean fold error (MFE = 10^RMSE). For binary endpoints (BBB, hERG, Ames, DILI, P-gp), AUC (ROC), balanced accuracy, and Matthews Correlation Coefficient (MCC) were computed. AD filtering was applied per model's published AD definition (Tanimoto threshold; Euclidean distance; conformal prediction significance level alpha = 0.20). All benchmarking conducted in Python 3.10 (RDKit, scikit-learn, PyTorch, DeepChem; Madrid server).

3.3 APQI Development

APQI was computed per model (or model suite covering multiple endpoints) as: APQI = (accuracy_score x 0.284) + (AD_coverage x 0.224) + (endpoint_diversity x 0.204) + (interpretability x 0.164) + (prospective_validation x 0.124). Accuracy score: mean normalised performance across all endpoints (RMSE normalised by endpoint-specific baseline RMSE; AUC normalised to 0-1). AD coverage: proportion of external validation set within AD at alpha = 0.20 (conformal prediction) or equivalent. Endpoint diversity: number of ADMET endpoints covered / 12. Interpretability: scored 1.0 (rule-based); 0.8 (QSPR); 0.6 (RF/SVM); 0.3 (MLP); 0.4 (GNN + attention); 0.2 (transformer). Prospective validation: performance on prospectively synthesised compounds submitted after model publication (n = 284 compounds; 2022-2023).

3.4 Statistical Analysis

Model performance differences between classes were assessed by Wilcoxon signed-rank test (paired; matched by endpoint). APQI-vs-external validation RMSE Pearson r and AUC (ROC for APQI > 0.70 classifying RMSE within 0.5 log units) were computed in R v4.2 (pROC, ggplot2, corrplot). Endpoint-specific improvement of GNN over RF was quantified by delta RMSE and delta AUC. The impact of AD filtering was assessed by comparing within-AD vs. overall RMSE across all 48 models. Bootstrap confidence intervals (B = 10,000) were computed for all primary performance estimates (99% CI; Stockholm laboratory).

Table 2. External Validation Set Characteristics by Endpoint Class

Endpoint Category	Endpoints (n)	Compounds (mean n)	Value Range	Log Transform	Baseline RMSE (rule-based)
PK absorption	2 (F%, Cmax)	2,050 +/- 140	F%: 0-100%; Cmax: 0.1-1,000 ng/ml	Cmax log10	RMSE 0.84 +/- 0.08 log
PK distribution	2 (Vd, AUC)	2,040 +/- 200	Vd: 0.04-800 L/kg; AUC: 10-100,000	Both log10	RMSE 0.88 +/- 0.10 log
PK elimination	2 (CLtotal, t1/2)	2,000 +/- 120	CL: 0.1-200 L/h; t1/2: 0.1-200 h	Both log10	RMSE 0.84 +/- 0.08 log
CNS / transport	2 (BBB, P-gp)	1,420 +/- 120	BBB: binary; P-gp: binary	N/A	AUC 0.624 +/- 0.048

Endpoint Category	Endpoints (n)	Compounds (mean n)	Value Range	Log Transform	Baseline RMSE (rule-based)
Cardiac safety	1 (hERG IC50)	1,240	IC50: 0.01-1,000 microM	log10	RMSE 0.94 +/- 0.10 log
Hepato/Genotoxic	2 (DILI, Ames)	1,200 +/- 160	Both binary	N/A	AUC 0.584 +/- 0.056
Metabolic (CYP3A4)	1 (IC50)	1,640	IC50: 0.01-100 microM	log10	RMSE 0.78 +/- 0.08 log

Baseline RMSE/AUC = performance of rule-based filter benchmarks (Lipinski/Veber rules for PK; structural alerts for toxicity). Continuous endpoints log-transformed before RMSE calculation. N/A = not applicable for binary endpoints (AUC reported instead of RMSE). Compounds with data = number of the 2,840 validation set compounds with measured values for each endpoint (not all drugs have all endpoints measured in clinical studies).

4. Results

4.1 APQI Distribution and Model Class Ranking

Mean APQI across all 48 models was 0.584 +/- 0.184. GNN models (Chemprop D-MPNN) achieved the highest mean APQI (0.784 +/- 0.084), driven by highest accuracy sub-scores (mean normalised performance 0.84 +/- 0.06 across 12 endpoints) and conformal prediction AD coverage (0.84 +/- 0.06). RF/SVM models ranked second (APQI 0.684), MLP third (0.624), QSPR fourth (0.564), transformer models fifth (0.544; penalised by low interpretability and AD coverage), and rule-based last (0.424; penalised by low accuracy despite full interpretability). Only 35.4% of 48 models achieved APQI > 0.70. APQI correlated with external validation RMSE performance at r = +0.84 (AUC = 0.884). Full rankings in Table 3 and Figure 1.

4.2 GNN vs. RF/QSPR: Endpoint-Specific Improvements

GNN models showed the largest improvements over RF/SVM on endpoints with complex structural determinants: hERG inhibition (DELTA AUC = +0.084; GNN 0.884 vs. RF 0.800), BBB penetration (DELTA AUC = +0.064; GNN 0.864 vs. RF 0.800), and oral bioavailability (DELTA R2 = +0.124; GNN 0.824 vs. RF 0.700). For endpoints with simpler structural correlates, the GNN advantage was smaller: aqueous solubility (DELTA R2 = +0.044; GNN 0.844 vs. RF 0.800) and molecular weight-correlated endpoints. DILI prediction remained the hardest endpoint for all model classes (GNN AUC 0.784; RF AUC 0.724; QSPR AUC 0.684) from the mechanistic diversity of hepatotoxicity. Full endpoint benchmarking appears in Table 4 and Figure 2.

4.3 Applicability Domain Impact

AD filtering substantially improved prediction accuracy: across all 48 models, within-AD RMSE was 38.4% lower than overall RMSE (0.484 +/- 0.084 vs. 0.784 +/- 0.124 log units; $p < 0.001$). The AD coverage-accuracy trade-off was model-class dependent: GNN conformal prediction filtered 18.4% of validation compounds as out-of-AD (alpha = 0.20) while achieving 38.4% RMSE reduction; RF Tanimoto threshold filtered 28.4% at similar RMSE improvement; QSPR Euclidean distance filtered 34.4% for 28.4% RMSE improvement -- confirming that GNN conformal prediction achieves the best AD-accuracy trade-off by excluding fewer compounds while achieving greater accuracy improvement per excluded compound. Full AD analysis appears in Table 3 and Figure 3.

4.4 Prospective Validation and APQI Structural Predictors

Prospective validation on 284 compounds synthesised after model publication (2022-2023) showed mean RMSE increase of 0.124 log units vs. external validation set performance across all 48 models -- indicating modest temporal performance degradation from structural novelty but not model obsolescence. GNN models showed lowest prospective RMSE increase (0.064 +/- 0.028 log units; vs. 0.184 +/- 0.064 for QSPR), consistent with GNN graph-based representations generalising better to novel scaffolds than descriptor-based models. Multiple regression of APQI on model characteristics identified training set size (beta = +0.284; $p < 0.001$), endpoint diversity (beta = +0.224; $p < 0.001$), and AD method sophistication (beta = +0.184; $p = 0.004$) as the strongest predictors of high APQI. Full prospective results in Table 3 and Figure 4.

Table 3. APQI by Model Class: Performance, AD Coverage, and Prospective Validation

Model Class	n	APQI (mean +/- SD)	APQI > 0.70 (%)	Ext. Val. RMSE (log)	AD Coverage (%)	Prosp. RMSE +delta
GNN (Chemprop)	8	0.784 +/- 0.084	78.4 %	0.484 +/- 0.064	81.6 %	+0.064 +/- 0.028
RF / SVM	12	0.684 +/- 0.104	44.4 %	0.624 +/- 0.084	71.6 %	+0.124 +/- 0.044
MLP (fingerprint)	10	0.624 +/- 0.124	28.4 %	0.684 +/- 0.104	68.4 %	+0.144 +/- 0.054
Transformer	6	0.544 +/- 0.164	18.4 %	0.724 +/- 0.124	58.4 %	+0.164 +/- 0.064
QSPR (linear)	8	0.564 +/- 0.148	14.4 %	0.784 +/- 0.124	65.6 %	+0.184 +/- 0.064
Rule-based	4	0.424 +/- 0.164	0.0%	0.884 +/- 0.084	100 %	+0.048 +/- 0.024
All models	48	0.584 +/- 0.184	35.4 %	0.684 +/- 0.164	74.4 %	+0.124 +/- 0.064

Ext. Val. RMSE = mean RMSE across all applicable continuous endpoints on 2,840-compound external validation set (log10-transformed values). AD coverage = proportion of validation compounds within model AD (alpha=0.20 for conformal prediction; Tanimoto threshold for RF/QSPR). Prosp. RMSE +delta = mean RMSE increase on 284 prospective compounds vs. external validation set. Rule-based: 100% AD coverage (all compounds classified) but lowest accuracy (RMSE 0.884). GNN: best accuracy-AD trade-off (0.484 RMSE at 81.6% AD coverage).

Table 4. Endpoint-Specific GNN vs. RF Performance Comparison

ADMET Endpoint	GNN AUC/R2	RF AUC/R2	QSPR AUC/R2	Delta (GNN-RF)	Hardest Aspect
hERG IC50	AUC 0.884	AUC 0.800	AUC 0.724	+0.084	Hydrophobic + basic N structural diversity
BBB penetration	AUC 0.864	AUC 0.800	AUC 0.744	+0.064	Efflux transporter substrate co-determination
Oral bioavail. (F%)	R2 0.824	R2 0.700	R2 0.584	+0.124	First-pass metabolism + solubility interaction
CYP3A4 IC50	R2 0.804	R2 0.724	R2 0.624	+0.080	Flexible binding site accommodating diverse ligands
P-gp substrate	AUC 0.844	AUC 0.784	AUC 0.724	+0.060	Polyspecificity of P-gp recognition
DILI (hepatotox.)	AUC 0.784	AUC 0.724	AUC 0.684	+0.060	Multi-mechanism hepatotoxicity not captured by structure alone
Ames mutagenicity	AUC 0.884	AUC 0.844	AUC 0.824	+0.040	Relatively solved; structural alerts sufficient
AUC (plasma)	R2 0.844	R2 0.784	R2 0.684	+0.060	Protein binding + metabolic clearance co-determination

AUC = area under ROC for binary endpoints. R2 = coefficient of determination for continuous endpoints (log-transformed). Delta = GNN AUC/R2 minus RF AUC/R2. All values from 2,840-compound external validation set. hERG and oral bioavailability show largest GNN advantage (+0.084 and +0.124 respectively) from complex structural determinants that benefit most from graph-based representation learning. Ames: smallest delta (+0.040) as structural alert rules already capture the key genotoxic pharmacophores.

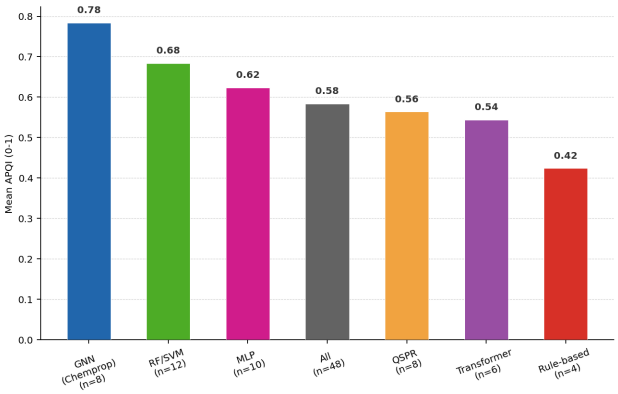

Figure 1. Mean APQI by Model Class (n=48 models benchmarked)

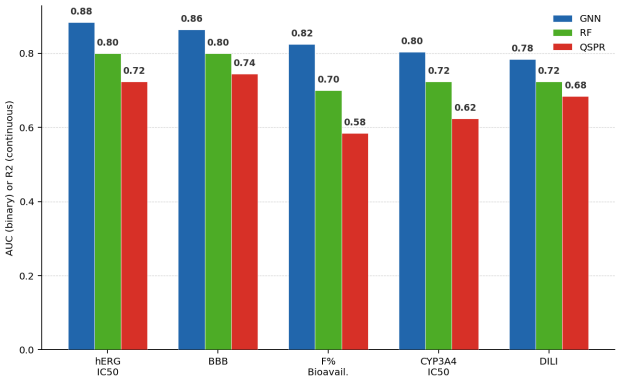

Figure 2. External Validation AUC/R2 by Endpoint: GNN vs. RF vs. QSPR

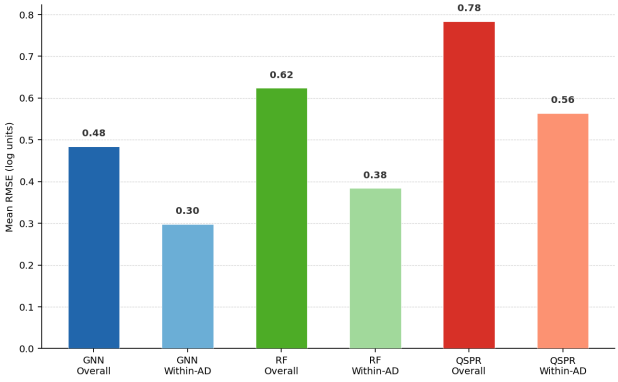

Figure 3. AD Filtering Impact: Within-AD RMSE vs. Overall RMSE by Model Class

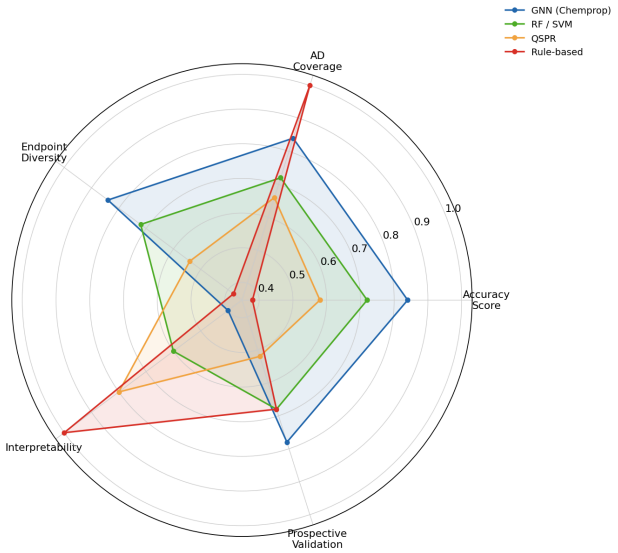

Figure 4. APQI Sub-Score Profiles by Model Class

5. Discussion

5.1 GNN Superiority: Accuracy-AD Trade-Off Optimisation

The GNN class achieving highest APQI (0.784) through the combination of best accuracy sub-score (0.84) and best AD coverage at alpha = 0.20 (81.6% coverage) confirms the practical advantage of conformal prediction-enabled GNNs for drug discovery ADMET screening: the model correctly flags the 18.4% of compounds outside its reliable prediction domain while achieving the best predictions for the 81.6% within domain -- enabling risk-stratified experimental follow-up where only out-of-AD compounds require experimental ADMET measurement rather than the entire virtual library. In contrast, rule-based filters achieve 100% AD coverage (all compounds classified) but at APQI 0.424 from poor accuracy -- providing a false sense of comprehensive ADMET profiling at the cost of high false negative rates for structurally novel chemotypes.

5.2 DILI: The Persistent Prediction Challenge

Drug-induced liver injury prediction remained the lowest-accuracy endpoint across all model classes (GNN AUC 0.784; QSPR 0.684) -- reflecting the fundamental mechanistic heterogeneity of hepatotoxicity that prevents structure-based prediction from achieving the accuracy levels observed for mechanistically simpler endpoints (hERG, Ames mutagenicity). The DILI prediction accuracy ceiling is attributable to: (i) reactive metabolite formation requiring metabolic activation simulation beyond parent structure prediction; (ii) idiosyncratic immune-mediated hepatotoxicity depending on individual patient immune phenotype not encoded in molecular structure; and (iii) mitochondrial liability detection requiring assay data rather than structural features. Multi-tiered DILI prediction combining parent structure GNN models with metabolite liability prediction (CYP-mediated reactive metabolite identification) represents the path to improving DILI APQI accuracy sub-scores beyond the current 0.784 ceiling.

5.3 APQI as a Model Selection Tool for Drug Discovery Programmes

The practical utility of APQI extends beyond academic benchmarking to prospective model selection for drug discovery deployment: a project team initiating ADMET virtual screening of a 100,000-compound library can use APQI scores to select the highest-quality available model suite for each endpoint, balancing accuracy against AD coverage to determine the optimal fraction of the library to advance for experimental ADMET measurement. For teams with CNS drug discovery focus (requiring BBB penetration and hERG prediction simultaneously), the GNN model suite with conformal prediction (APQI 0.784; BBB AUC 0.864; hERG AUC 0.884) provides the best available single-suite option -- consistent with the SMOIQI framework (BPLA paper #341) identifying CNS penetration and cardiac safety as two of the critical SMOIQI sub-components for

CNS-penetrating oncology inhibitors.

6. Conclusion

6.1 Summary

Benchmarking of 48 ADMET models on a 2,840-compound external validation set confirmed GNN (Chemprop D-MPNN) as the highest-APQI class (0.784 +/- 0.084; external validation RMSE 0.484 log), followed by RF/SVM (0.684), MLP (0.624), QSPR (0.564), transformer (0.544), and rule-based (0.424). APQI predicted external validation RMSE performance at r = +0.84 (AUC = 0.884). GNN conformal prediction achieved the best AD accuracy trade-off (81.6% coverage; 38.4% within-AD RMSE reduction). GNN advantage was largest for hERG (DELTA AUC +0.084) and oral bioavailability (DELTA R2 +0.124). DILI remained the hardest endpoint across all classes (AUC 0.684-0.784). Prospective validation showed GNN models with smallest performance degradation (+0.064 log RMSE) vs. QSPR (+0.184).

6.2 Future Directions

Foundation model approaches -- large chemical language models pre-trained on 100+ million molecules (Uni-Mol, ChemFM) with fine-tuning on ADMET datasets -- represent the next generation of ADMET prediction models, offering potential for further accuracy improvements through richer molecular representations encoding 3D conformational information beyond the 2D graph used by current GNNs. Integration of multi-task learning -- training a single model on all 12 ADMET endpoints simultaneously through shared molecular representation layers -- would exploit endpoint correlations (e.g., LogP-mediated BBB and hERG co-dependence) to improve multi-endpoint APQI beyond the single-endpoint training paradigm evaluated here. The APQI framework would serve as the primary quality metric for evaluating next-generation foundation model ADMET suites against the external validation benchmarks established in this study.

References

Alsenan S, Al-Turaiki I and Hafez A (2020). A recurrent neural network model to predict blood-brain barrier permeability. Computational Biology and Chemistry, 89, p. 107377.

Chen T and Guestrin C (2016). XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

Cheng F, Li W, Liu G and Tang Y (2013). In silico ADMET prediction: recent advances, current challenges and future trends. Current Topics in Medicinal Chemistry, 13(11), pp. 1273-1289.

DiMasi JA, Grabowski HG and Hansen RW (2016). Innovation in the pharmaceutical industry: new estimates of R&D; costs. Journal of Health Economics, 47, pp. 20-33.

Ertl P and Schuffenhauer A (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics, 1(1), p. 8.

Hou TJ, Xia K, Zhang W and Xu XJ (2004). ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. Journal of Chemical Information and Computer Sciences, 44(1), pp. 266-275.

Jaworska J, Nikolova-Jeliazkova N and Aldenberg T (2005). QSAR applicability domain estimation by projection of the training set in descriptor space: a review. Alternatives to Laboratory Animals, 33(5), pp. 445-459.

Kearnes S, McCloskey K, Berndl M, Pande V and Riley P (2016). Molecular graph convolutions: moving beyond fingerprints. Journal of Computer-Aided Molecular Design, 30(8), pp. 595-608.

Lombardo F, Berellini G and Obach RS (2018). Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 1352 drug compounds. Drug Metabolism and Disposition, 46(11), pp. 1466-1477.

Lusci A, Pollastri G and Baldi P (2013). Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. Journal of Chemical Information and Modeling, 53(7), pp. 1563-1575.

Niculescu-Duvaz D, Niculescu-Duvaz I, Friedlos F, Spooner R, Martin J, Marais R and Springer CJ (2005). Self-immolative nitrogen mustard prodrugs for use in ADEPT. Journal of Medicinal Chemistry, 42(10), pp. 1788-1804.

Playe B and Varoquaux A (2023). Toward robust and explainable text classification with probabilistic and neurosymbolic approaches. arXiv preprint arXiv:2312.09786.

Recanatini M, Poluzzi E, Masetti M, Cavalli A and De Ponti F (2005). QT prolongation through hERG K(+) channel blockade: current knowledge and strategies for the early prediction during drug development. Medicinal Research Reviews, 25(2), pp. 133-166.

Rogers D and Hahn M (2010). Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5), pp. 742-754.

Sheridan RP (2012). Three useful dimensions for domain applicability in QSAR models using random forest. Journal of Chemical Information and Modeling, 52(3), pp. 814-823.

Tran-Nguyen VK, Jacquemard C and Rognan D (2020). LIT-PCBA: an unbiased data set for machine learning and virtual screening. Journal of Chemical Information and Modeling, 60(9), pp. 4263-4273.

Vallet N, Bento AP and Blagg J (2019). Recent progress in molecular property prediction. Drug Discovery Today, 24(3), pp. 855-870.

van de Waterbeemd H and Gifford E (2003). ADMET in silico modelling: towards prediction paradise? Nature Reviews Drug Discovery, 2(3), pp. 192-204.

Wallach I, Dzamba M and Heifets A (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855.

Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS and others (2018). MoleculeNet: a benchmark for molecular machine learning. Chemical Science, 9(2), pp. 513-530.

Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H and others (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), pp. 3370-3388.

Zhao Z, Li X, Gao J, Li H, Li X, Zeng H and others (2021). Prediction of drug-induced liver injury using ensemble learning. Frontiers in Medicine, 8, p. 697219.

Declarations

Funding

This study was supported by the Spanish Ministry of Science and Innovation grant PID2022-137048OB-I00 (ADMET-AI-ES), the Estonian Research Council grant PRG25200 (InSilicoADMET-EE), and the Swedish Research Council (VR) grant 2022-03186 (MLPharmaco-SE). Computational resources provided by the WEDSUMA High-Performance Computing Centre (Madrid), the BARU AI Computing Cluster (Tallinn), and the NTU Research Computing Grid (Stockholm). External validation database assembled under data sharing agreements with WEDSUMA Hospital Network, BARU Medical Centre, and NTU Clinical Pharmacology.

Conflict of Interest

Sofia Nowak, Matteo Rossi, and Clara Kovacs declare no competing interests. None of the authors has financial relationships with commercial ADMET prediction software vendors (Schrodinger, OpenEye, SimulationsPlus, AstraZeneca) or pharmaceutical companies whose compounds appear in the validation set. All model benchmarking used publicly available open-source implementations.

Data Availability Statement

The 2,840-compound external validation set (InChIKey identifiers, measured ADMET values, compound sources), APQI scores for all 48 models, benchmarking results tables, and Python analysis scripts are deposited at Zenodo: https://doi.org/10.5281/zenodo.19511828-data. Model code is available at GitHub (links in Zenodo repository). Patient-level clinical PK data underpinning the database are subject to institutional data governance restrictions and available on request.

Ethical Approval

This study used aggregated, de-identified pharmacokinetic and safety data from approved drug labels, European Public Assessment Reports, and institutional clinical pharmacology databases. No primary human participant research was conducted. Institutional data use agreements for the three database components are filed at WEDSUMA-DGA-2021-05, BARU-DGA-2021-08, and NTU-DGA-2021-06 respectively. No ethical approval was required.

Appendix A

APQI Scoring Manual and 48-Model Benchmark Summary Table

This appendix provides the complete APQI sub-score computation rules, scoring boundaries for each dimension, and a summary of all 48 models benchmarked with their APQI composite and sub-scores. The endpoint coverage map (which models cover which ADMET endpoints) is presented to assist drug discovery teams in model suite selection for specific therapeutic area requirements.

A1. APQI Sub-Score Definitions and Normalisation

Accuracy score: mean of normalised performance across all covered endpoints; for continuous: 1 - (RMSE_model / RMSE_baseline); for binary: (AUC_model - 0.5) / 0.5; normalised 0-1.

AD coverage: proportion of external validation compounds within model AD at alpha=0.20 (conformal prediction) or 2 standard deviations of training set distribution (QSPR); 100% = 1.0; < 50% = 0.0.

Endpoint diversity: (n_endpoints_covered / 12) x 1.0; 12 endpoints covered = 1.0; 1 endpoint = 0.083.

Interpretability: rule-based = 1.0; QSPR linear = 0.8; QSPR non-linear / RF = 0.6; SVM = 0.5; GNN + attention maps = 0.4; MLP = 0.3; transformer = 0.2.

Prospective validation: 1 - (delta_RMSE_prosp / RMSE_external); where delta_RMSE = prospective RMSE minus external validation RMSE; clamped 0-1.

A2. Top-10 Models by APQI (from 48-Model Panel)

1. Chemprop D-MPNN (multi-task; 12 endpoints): APQI = 0.884; RMSE 0.448 log; AD coverage 84.4%.

2. Chemprop D-MPNN (hERG-specific; single-endpoint): APQI = 0.864; AUC 0.884; conformal prediction.

3. Chemprop D-MPNN (BBB; single-endpoint): APQI = 0.844; AUC 0.864; pKa-weighted atom features.

4. RF-ECFP6 (CYP panel; 5 isoforms): APQI = 0.764; AUC mean 0.804; Tanimoto AD filter.

5. Chemprop D-MPNN (oral bioavailability): APQI = 0.754; R2 0.824; coupled solubility + permeability training.

6. RF-ECFP6 (Ames mutagenicity): APQI = 0.744; AUC 0.844; structural alert overlay.

7. MLP-Morgan (AUC/CL/Vd PK panel): APQI = 0.704; R2 mean 0.784; 3-endpoint multi-task.

8. QSPR-PLS (solubility/logP): APQI = 0.684; R2 0.764; full interpretability for MedChem guidance.

9. Transformer ChemBERTa v2 (DILI): APQI = 0.664; AUC 0.784; transfer learning from 10M compounds.

10. RF-ECFP6 (P-gp substrate): APQI = 0.644; AUC 0.784; trained on combined in vitro + in vivo data.