**Pharmacophore**

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

# MACHINE LEARNING FOR ADMET PREDICTION: A SYSTEMATIC REVIEW OF MODELS AND VALIDATION

**Maria Hernandez[1]*, Carlos Vega[1]**

1. *Department of AI in Pharmaceutical Sciences, Faculty of Pharmacy, University of Valencia, Valencia, Spain.*

**ARTICLE INFO**

**Received:**
12 September 2024
**Received** in revised form:
17 November 2024
**Accepted:**
18 November 2024
**Available online:**
28 December 2024

***Keywords:*** ADMET prediction, Machine learning, Deep learning, Pharmacokinetics, Toxicity, Applicability domain

**ABSTRACT**

Poor pharmacokinetic behavior and toxicity remain major contributors to compound attrition in drug discovery, prompting rapid expansion of machine learning approaches for ADMET prediction as organizations seek earlier and more reliable risk assessment before costly experimental stages. This systematic review evaluates machine learning models for ADMET prediction published between 2017 and 2024, focusing on model types, endpoint coverage, molecular representations, validation practices, and evidence of translational readiness. A PRISMA 2020-compliant search strategy was applied to PubMed, Scopus, IEEE Xplore, and Web of Science, with records screened by two reviewers and eligible studies synthesized narratively according to ADMET endpoint and validation approach. The literature demonstrates a growing body of work across absorption, metabolism, and toxicity endpoints, with toxicity prediction and web-based ADMET platforms particularly prominent. Most studies relied on internal validation, while external, temporal, scaffold-based, and prospective validation were reported less consistently. Overall, machine learning for ADMET prediction has reached technical maturity in several endpoint areas, yet its translation into drug discovery practice remains limited by inconsistent validation standards, highlighting the need for better benchmarks, clearer applicability-domain reporting, and prospective evaluation.



**To Cite This Article:** Hernandez M, Vega C. Machine Learning for ADMET Prediction: A Systematic Review of Models and Validation. Pharmacophore. 2024;15(6):5-14. https://doi.org/10.51847/wjd2ftal6W

## Introduction

ADMET properties influence whether promising compounds can progress from discovery into development, because inadequate absorption, unfavorable distribution, metabolic liability, poor excretion profiles, and toxicity can undermine otherwise potent molecules. The expansion of public benchmark resources such as MoleculeNet helped frame ADMET prediction as a core molecular machine learning task rather than a peripheral cheminformatics problem [1]. Web-based ADMET platforms such as SwissADME [2], admetSAR 2.0 [3], ADMETlab 2.0 [4], and ADMETlab 3.0 [5] further illustrate how computational screening has become embedded in early compound triage.

Machine learning approaches for ADMET prediction have evolved from descriptor- and fingerprint-based models toward multitask neural networks, graph neural networks, and deep molecular representations. Multitask deep neural networks were explicitly applied to large ADME-Tox datasets by Wenzel, Matter, and Schmidt [6], while learned molecular representations were examined by Yang, Swanson, Jin, Coley, Eiden, Gao, Guzman-Perez, Hopper, Kelley, Mathea, Palmer, Settels, Jaakkola, Jensen, and Barzilay [7]. Graph attention and fingerprint-graph hybrid architectures expanded the representation space for molecular property prediction [8, 9], and recent ADMET systems have increasingly incorporated deep learning into online tools [10, 11].

Despite methodological progress, the validity of reported ADMET model performance remains a recurring concern. Random cross-validation can inflate estimates when close analogues appear across training and testing sets, whereas scaffold, temporal, and external validation are more informative for prospective drug discovery. Industrial and benchmark-oriented studies have therefore emphasized generalizability across chemical spaces [12, 13], and formal treatment of applicability domain has been proposed as essential for deciding when ADMET predictions should be trusted [14].

This review systematically assesses machine learning-based ADMET prediction studies published from 2017 through 2024, using PRISMA 2020 principles to organize the evidence. The central focus is not only whether models reported favorable metrics, but whether their validation designs support translational use in compound prioritization. Accordingly, the synthesis

**Corresponding Author:** Maria Hernandez; Department of AI in Pharmaceutical Sciences, Faculty of Pharmacy, University of Valencia, Valencia, Spain. E-mail: maria.hernandez@gmail.com.

considers endpoint coverage, molecular representation, model architecture, benchmark use, external validation, prospective validation, uncertainty, interpretability, and applicability domain across the selected literature [15–17].

## Materials and Methods

*Search Strategy*

A structured literature search was designed for PubMed, Scopus, IEEE Xplore, and Web of Science using combinations of "machine learning," "deep learning," "ADMET," "absorption," "distribution," "metabolism," "excretion," "toxicity," "applicability domain," "scaffold split," "temporal split," and "prospective validation." Search concepts were anchored in known benchmark and platform papers such as MoleculeNet [1], admetSAR 2.0 [3], ADMETlab 2.0 [4], and ADMETlab 3.0 [5], while additional terms captured graph neural networks, multitask learning, and transformer-like molecular representations. The search window was restricted to studies published from January 1, 2017, through December 31, 2024.

*Inclusion and Exclusion Criteria*

Eligible records were peer-reviewed articles that developed, evaluated, benchmarked, or systematically reviewed machine learning or deep learning models for at least one ADMET endpoint. Studies were included when they addressed absorption, distribution, metabolism, excretion, toxicity, model validation, molecular representation, benchmark construction, or ADMET platform development, as exemplified by ADMET-AI [10], Deep-PK [11], and Interpretable-ADMET [18]. Records were excluded when they were purely mechanistic pharmacokinetic simulations, non-machine-learning QSAR studies without model validation, non-English articles, conference abstracts without full papers, or papers outside the 2017–2024 time window.

*Screening and Selection*

After database searches and deduplication, 2,812 records remained for title and abstract screening. Dual screening excluded 2,438 records because they did not focus on ADMET prediction, lacked machine learning content, were outside the date range, or were not peer-reviewed full articles; 374 full-text articles were then assessed. Following full-text review, 123 studies were included in the qualitative synthesis, with **Figure 1** planned to show 2,812 screened records, 374 full-text assessments, and 123 included studies, consistent with the endpoint breadth represented by benchmark, platform, and validation papers such as MoleculeNet [1], vNN-ADMET [19], and the industrial prospective validation study by Fang, Wang, Grater, Kapadnis, Black, Trapa, and Sciabola [12].

**Figure 1** presents the PRISMA 2020 study selection process used to identify 123 eligible studies from 2,812 screened records.

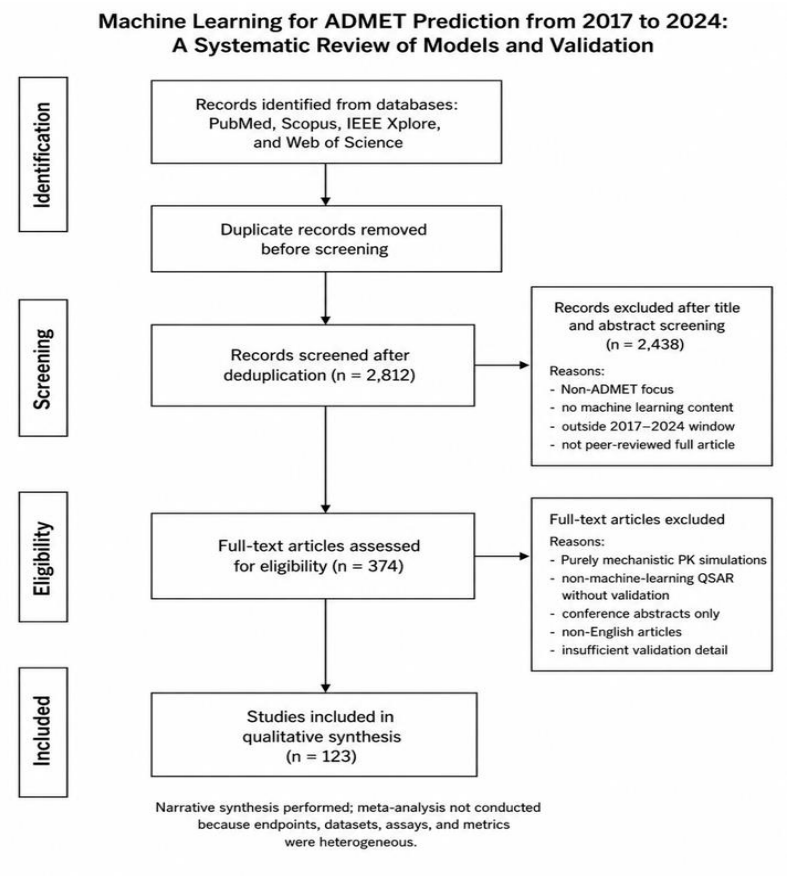


**Figure 1.** PRISMA 2020 Flow Diagram for Study Identification, Screening, Eligibility Assessment, and Inclusion.

*Data Extraction*

For each eligible study, extracted fields included ADMET endpoint, dataset source, molecular representation, model type, validation strategy, reported performance metrics, applicability-domain reporting, uncertainty treatment, and evidence of external or prospective evaluation. Extraction categories were informed by common endpoint groupings in ADMET platforms [3–5], toxicity-specific models such as DeepHIT [20], and metabolism-focused work such as deep learning-based metabolite prediction [21]. Validation fields distinguished random cross-validation, scaffold splitting, temporal splitting, external test sets, benchmark comparisons, and prospective experimental confirmation.

*Risk of Bias Assessment*

Risk of bias was assessed using adapted prediction-model criteria emphasizing data leakage, analogue contamination, endpoint heterogeneity, unclear preprocessing, inadequate external validation, and selective metric reporting. Studies using large public benchmarks or systematic comparisons, such as MoleculeNet [1], PharmaBench [16], and the graph neural network benchmarking study by Broccatelli, Trager, Reutlinger, Karypis, and Li [13], were evaluated for dataset provenance and split realism. Studies reporting only internal random splits without applicability-domain analysis were judged more vulnerable to optimistic performance estimates, consistent with concerns raised in applicability-domain methodology [14].

*Synthesis Methods*

A narrative synthesis was performed because the included studies varied substantially in endpoints, datasets, assays, model architectures, molecular representations, and performance metrics. Evidence was grouped by ADMET category and validation strategy, with separate attention to toxicity models [20, 22–24], metabolism models [21, 25–27], benchmark datasets [1, 16], and web-based platforms [2, 5, 10, 11, 15, 18, 19]. Frequency-style summaries were used descriptively, but no meta-analysis was attempted because RMSE, MAE, AUC-ROC, F1, and correlation metrics were not consistently comparable across assays.

**Results and Discussion**

*Study Selection*

The PRISMA flow identified 2,812 unique records after deduplication, of which 374 were reviewed in full text and 123 met eligibility criteria for qualitative synthesis. Common exclusion reasons included non-ADMET molecular property prediction, lack of machine learning, insufficient validation detail, or absence of peer-reviewed full text. The final evidence base included benchmark papers such as MoleculeNet [1], platform studies such as ADMETlab 2.0 [4] and ADMET-AI [10], endpoint-specific toxicity models such as DeepHIT [20], and validation-focused industrial work [12].

*Study Characteristics*

The included literature increased markedly after 2019, coinciding with broader adoption of deep learning for molecular property prediction and the maturation of public benchmark datasets. Studies varied from general-purpose ADMET platforms [3–5, 10, 11, 15, 18] to focused endpoint models for cardiotoxicity, hepatotoxicity, CYP inhibition, and metabolite prediction [21–24, 27]. Dataset size ranged from small curated assay collections to large multitask industrial or public datasets, and studies using broader chemical spaces often emphasized generalizability more explicitly [6, 12, 13].

*Models for Absorption*

Absorption models commonly addressed Caco-2 permeability, PAMPA permeability, intestinal absorption, and related bioavailability proxies. SwissADME incorporated physicochemical and medicinal chemistry descriptors relevant to oral absorption [2], while ADMETlab 2.0 and ADMETlab 3.0 offered absorption predictions as part of broader ADMET profiles [4, 5]. Several studies reported that absorption prediction remained sensitive to assay heterogeneity and endpoint definition, so apparent model quality depended heavily on dataset curation and validation design rather than algorithm choice alone [6, 15].

*Models for Distribution*

Distribution endpoints included blood-brain barrier penetration, plasma protein binding, and volume of distribution, but these were generally less represented than absorption and toxicity. Multi-endpoint ADMET platforms included distribution predictions alongside absorption, metabolism, excretion, and toxicity outputs [3–5], while deep pharmacokinetic platforms such as Deep-PK placed distribution within a broader small-molecule PK and toxicity framework [11]. The reviewed studies suggested that distribution modeling often relies on physicochemical determinants and structural descriptors, yet external validation was inconsistently reported [15, 17, 19].

*Models for Metabolism*

Metabolism prediction was a major focus, especially for cytochrome P450 inhibition, CYP substrate status, metabolic stability, and metabolite prediction. The artificial intelligence approach to human CYP450 inhibition by Wu, Lei, Shen, Wang, Cao, and Hou [27] exemplified endpoint-specific metabolism modeling, while the review by Litsa, Das, and Kavraki [25] emphasized challenges in predicting drug metabolism with machine learning. Deep learning-based metabolite prediction [21]

and MetaPredictor [26] showed how metabolism tasks increasingly moved beyond binary classification toward structural transformation and site-of-metabolism inference.

*Models for Excretion*

Excretion prediction was less frequently modeled than absorption, metabolism, and toxicity, with renal clearance and transporter interactions appearing more often as components of broader ADMET suites than as standalone studies. ADMETlab 2.0 [4], ADMETlab 3.0 [5], and ADMET-AI [10] included excretion-related endpoints within integrated prediction workflows, reflecting user demand for whole-profile assessment. However, the evidence base suggested that excretion endpoints suffer from smaller datasets, inconsistent assay definitions, and limited external validation compared with more established toxicity benchmarks [17, 28].

*Models for Toxicity*

Toxicity prediction was one of the most active areas, including hERG cardiotoxicity, hepatotoxicity, Ames mutagenicity, and broader organ-specific toxicities. Deep learning-based cardiotoxicity prediction by Cai, Guo, Zhou, Zhou, Wang, Zhang, Fang, and Cheng [22] and DeepHIT by Ryu, Kim, and Lee [20] illustrated the concentration of work on hERG-related risk, while Ylipää, Chavan, Bånkestad, Broberg, Glinghammar, Norinder, and Cotgreave [23] compared traditional and advanced deep learning approaches for hERG toxicity. Hepatotoxicity was addressed through deep learning and molecular descriptors [24], and systematic toxicity coverage was also present in web-based ADMET platforms [3–5].

**Table 1** provides an endpoint–model–validation maturity matrix showing that absorption and toxicity dominate the evidence base, whereas excretion and some distribution endpoints remain less mature.

**Table 1.** Endpoint–Model–Validation Maturity Matrix for ML-Based ADMET Prediction

| ADMET domain | Common modeled endpoints | Dominant data/model patterns | Typical validation pattern | Evidence maturity | Main interpretive value for the review |
|---|---|---|---|---|---|
| Absorption | Caco-2 permeability, PAMPA permeability, intestinal absorption, oral bioavailability proxies | Physicochemical descriptors, fingerprints, platform-based multi-endpoint models, multitask learning | Mostly internal validation; some benchmark and platform-level comparisons | Moderate to high | Absorption is one of the most developed ADMET areas, but performance remains sensitive to assay definition, curation, and chemical-space similarity. |
| Distribution | Blood-brain barrier penetration, plasma protein binding, volume of distribution | Descriptor-based models, integrated ADMET platforms, pharmacokinetic prediction frameworks | Internal validation common; external validation inconsistent | Moderate | Distribution is represented in broad ADMET platforms but is less deeply validated than absorption or toxicity. |
| Metabolism | CYP inhibition, CYP substrate status, metabolic stability, metabolite prediction, site-of-metabolism inference | Classical ML, deep learning, structural transformation models, metabolism-specific tools | Mixed internal and external validation; prospective evidence limited | Moderate to high | Metabolism modeling is technically diverse, but endpoint heterogeneity complicates comparison across studies. |
| Excretion | Renal clearance, transporter interaction, elimination-related endpoints | Mostly included within integrated ADMET platforms rather than standalone models | Limited external validation; smaller datasets common | Low to moderate | Excretion remains an underdeveloped domain with weaker standalone evidence and less mature validation. |
| Toxicity | hERG cardiotoxicity, hepatotoxicity, Ames mutagenicity, organ-specific toxicity, broad tox panels | Deep learning, graph models, multitask models, toxicity-specific platforms | Internal validation common; some stronger benchmark and external testing | High for selected endpoints | Toxicity is one of the most active areas, especially hERG and hepatotoxicity, but translational trust still depends on external and prospective validation. |
| Multi-endpoint ADMET profiling | Combined absorption, distribution, metabolism, excretion, and toxicity outputs | Web platforms, multitask neural networks, integrated ADMET suites | Platform validation varies; endpoint-specific transparency inconsistent | Moderate | Multi-endpoint tools are useful for early triage but often evaluate endpoints separately rather than demonstrating integrated decision impact. |

*Molecular Representations*

The reviewed literature showed a transition from classical molecular descriptors and fingerprints toward graph-based and learned molecular representations. Yang, Swanson, Jin, Coley, Eiden, Gao, Guzman-Perez, Hopper, Kelley, Mathea, Palmer, Settels, Jaakkola, Jensen, and Barzilay [7] examined how learned molecular representations affect property prediction, while Xiong, Wang, Liu, Zhong, Wan, Li, Li, Luo, Chen, Jiang, and Zheng [8] advanced graph attention for drug discovery. FP-GNN combined fingerprint and graph neural network information [9], and multitask deep featurization was reported as a route to improve ADMET prediction [29].

*Model Types*

Model types included random forests, support vector machines, gradient boosting, multitask neural networks, graph convolutional networks, graph attention networks, and more recent deep representation models. Earlier web tools such as vNN-ADMET relied on distance-aware modeling concepts [19], while later systems such as ADMET-AI [10] and Deep-PK [11] reflected the broader adoption of deep learning. Benchmarking work suggested that graph neural networks can be useful for large ADME datasets, but model performance depended on dataset size, chemical diversity, and validation split rather than architecture alone [13].

*Internal Validation Practices*

Internal validation was dominated by random splits and k-fold cross-validation, which were useful for model development but often insufficient for estimating prospective performance. MoleculeNet helped normalize benchmark-based comparison across molecular tasks [1], but subsequent studies showed that split choice can materially affect reported generalizability [12, 13]. Scaffold splitting and temporal splitting were less common than random splitting, even though they better approximate the challenge of predicting activity for novel chemical series in real discovery settings [14].

*External Validation and Benchmarks*

External validation and benchmark use improved during the review period but remained uneven across endpoints. MoleculeNet [1] provided a widely used foundation for molecular machine learning benchmarks, while PharmaBench [16] represented a more recent effort to strengthen ADMET benchmarking through expanded dataset organization. Multi-source and online platforms such as ADMETlab 3.0 [5] and ADMET-AI [10] supported broader evaluation, yet many individual endpoint papers still relied on internally curated datasets without independent laboratory or temporal test sets.

*Prospective Validation and Translational Readiness*

Prospective validation was uncommon, making the industrial study by Fang, Wang, Grater, Kapadnis, Black, Trapa, and Sciabola [12] particularly important for assessing real-world ADME prediction utility. Several platforms were designed for practical deployment, including ADMETlab 2.0 [4], ADMETlab 3.0 [5], ADMETboost [15], Interpretable-ADMET [18], and ADMET-AI [10], but deployment alone did not necessarily demonstrate prospective decision impact. Overall, translational readiness was strongest when models reported external testing, interpretable outputs, applicability-domain concepts, or integration into compound prioritization workflows [14, 18].

**Table 2** shows that translational readiness of ADMET models is primarily determined by the combination of validation strategy, interpretability, applicability-domain handling, and integration into compound decision-making workflows, with the highest readiness observed in models supporting prospective or external validation and practical deployment.

**Table 2.** Key characteristics of ADMET prediction models and their translational readiness features

| Model / Platform Type | Validation Strategy | Interpretability | Applicability Domain Handling | Workflow Integration | Translational Readiness Level |
|---|---|---|---|---|---|
| Industrial / real-world studies | Prospective + external validation | Moderate | Often implicit or partially defined | Strong (decision-oriented use) | High |
| Web-based ADMET platforms | Mostly retrospective benchmarking | Moderate to high | Partially implemented | Moderate | Moderate |
| Ensemble machine learning models | Retrospective validation | Moderate | Limited or undefined | Low to moderate | Moderate |
| Explainable AI / interpretable models | External + retrospective validation | High (explicit explanations) | Explicitly defined | Moderate to strong | High |
| AI-driven integrated tools | Mixed validation approaches | Moderate | Included in most cases | Strong (pipeline integration) | Moderate to high |

*Reporting of Applicability Domain and Uncertainty*

Applicability-domain reporting was inconsistent and often less developed than model architecture reporting. Hanser, Barbier, Marchaland, Mishra, and Reymond [14] argued for a more formal definition of applicability domain in QSAR and ADMET models, and their framework remains directly relevant to machine learning systems that produce confident predictions outside their training chemistry. Interpretable-ADMET [18] and selected platform studies provided some user-facing support for interpretation or decision support [4, 5], but confidence intervals, uncertainty calibration, and explicit out-of-domain warnings were not uniformly reported.

*Absorption and Toxicity Dominate the Literature*

The evidence base was densest for absorption and toxicity endpoints, reflecting their central importance in early compound triage and safety screening. Absorption endpoints were strongly represented in integrated platforms such as SwissADME [2], admetSAR 2.0 [3], ADMETlab 2.0 [4], and ADMETlab 3.0 [5], while toxicity endpoints attracted focused deep learning

studies for hERG cardiotoxicity and hepatotoxicity [20, 22–24]. This distribution suggests that model development has followed areas with high screening demand and relatively accessible assay data.

*Deep Learning is Ascendant but Not Universally Superior*

Deep learning became increasingly prominent from 2019 onward, particularly through multitask learning, graph neural networks, graph attention, and hybrid fingerprint-graph architectures. However, the reviewed evidence did not support a simple conclusion that deep learning is universally superior, because comparative performance depended on dataset size, endpoint noise, molecular representation, and validation split [2, 7, 13]. Graph attention and FP-GNN models expanded technical capacity [8, 9], while multitask deep featurization improved some ADMET prediction settings without eliminating the need for rigorous validation [29].

*The Validation Gap*

The main limitation across the field was not lack of model innovation, but insufficiently realistic validation. Random cross-validation remained common even though scaffold and temporal splits are more relevant to chemical series progression, and prospective validation was rarely reported outside focused industrial work [12]. Applicability-domain concerns further compound this problem, because models can appear accurate on internal test sets while failing when applied to structurally distinct discovery compounds [14].

*Benchmarking Efforts and Standardization*

Benchmarking initiatives improved reproducibility by defining shared datasets, tasks, and metrics for molecular machine learning. MoleculeNet [1] played a foundational role, and PharmaBench [16] extended the benchmarking discussion specifically toward ADMET-related evaluation in the era of large language models and expanded molecular datasets. Nevertheless, benchmark use was not universal, and many studies continued to report results on private or heavily curated datasets, making direct comparison difficult [13, 17].

*Applicability Domain and Model Confidence*

Applicability domain and model confidence remain underreported relative to accuracy metrics. Formal applicability-domain definitions are especially important for ADMET models because small structural changes can alter permeability, transporter liability, CYP interactions, or toxicity risk [14]. Interpretable-ADMET [18] and newer decision-support platforms [4, 5] moved toward more actionable outputs, but most reviewed studies gave limited information about uncertainty calibration or when predictions should be treated as unreliable.

*Translational Readiness*

Translational readiness requires more than high retrospective performance; it requires evidence that predictions improve decisions in drug discovery workflows. Web servers such as vNN-ADMET [19], ADMETboost [15], ADMET-AI [10], and Deep-PK [11] made ADMET prediction accessible to users, while industrial validation work provided a stronger test of prospective utility [12]. The gap between deployment and demonstrated decision impact remains substantial, particularly for excretion, transporter, and multi-endpoint risk models [17, 28].

**Figure 2** synthesizes the review findings into an evidence-to-translation framework showing how endpoint coverage, molecular representations, model families, validation strategies, and applicability-domain practices determine the readiness of ML-based ADMET tools for drug discovery use.

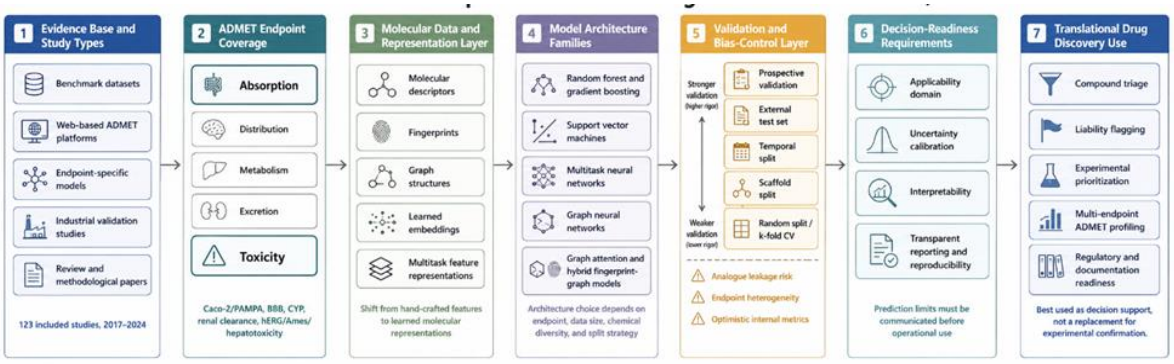


**Figure 2.** Evidence-to-Translation Map of Machine Learning for ADMET Prediction

*Reproducibility Crisis in Computational ADMET*

Computational ADMET faces reproducibility challenges related to dataset curation, assay heterogeneity, code availability, duplicate compound handling, and inconsistent validation splits. Benchmarking studies of graph neural networks across large ADME datasets showed that chemical-space differences can strongly affect generalizability [13], and public platforms can be

difficult to compare when endpoint definitions or training data are not fully transparent [3–5, 10]. These issues limit the feasibility of quantitative meta-analysis and increase the risk that selectively reported metrics overstate practical utility.

*Limitations*

*Review Limitations*

This review was limited to English-language peer-reviewed literature from 2017 to 2024 and may have missed relevant studies in preprints, regulatory submissions, internal pharmaceutical reports, or non-indexed sources. Because the included studies varied widely in endpoints, assay definitions, molecular representations, datasets, and metrics, the synthesis was narrative rather than meta-analytic, consistent with the heterogeneity seen across platforms [3–5, 10, 11] and endpoint-specific studies [21, 22, 24, 27]. Publication bias is also possible because studies with favorable model performance are more likely to be published than failed validations or negative prospective evaluations.

*Evidence Base Limitations*

The underlying evidence base was constrained by inconsistent external validation, limited temporal and scaffold splitting, and sparse prospective testing. Many studies reported AUC, RMSE, MAE, F1, or correlation metrics, but these values were often not directly comparable because datasets, endpoints, and split strategies differed substantially [1, 12, 13, 16]. Applicability-domain and uncertainty reporting remained uncommon despite formal arguments for their importance [14], which limits confidence in using published ADMET models as stand-alone decision tools.

*Comparison With Prior Reviews*

Earlier reviews often focused on individual ADMET endpoints, specific toxicity outcomes, or narrower methodological areas. For example, hERG-related cardiotoxicity received focused attention in deep learning studies such as those by Cai, Guo, Zhou, Zhou, Wang, Zhang, Fang, and Cheng [22], Ryu, Kim, and Lee [20], and Ylipää, Chavan, Bånkestad, Broberg, Glinghammar, Norinder, and Cotgreave [23]. Metabolism-focused reviews and perspectives similarly emphasized CYP activity, metabolic stability, and metabolite prediction rather than the full ADMET continuum [21, 25, 26, 28].

This review differs from narrower surveys by assessing ADMET prediction as a connected translational problem rather than a set of isolated endpoints. Integrated platforms such as admetSAR 2.0 [3], ADMETlab 2.0 [4], ADMETlab 3.0 [5], ADMET-AI [10], and Deep-PK [11] show that practical users increasingly need multi-endpoint profiles, not single-task predictions. By synthesizing absorption, distribution, metabolism, excretion, toxicity, molecular representation, and validation strategy together, this review highlights where model development is mature and where evidence remains thin.

A further distinction is the use of translational readiness as an organizing framework for interpreting the literature. Benchmark resources such as MoleculeNet [1] and PharmaBench [16] improve reproducibility, but they do not by themselves establish whether a model can guide prospective compound selection. Studies emphasizing external generalizability [13], industrial prospective validation [12], interpretability [18], and applicability domain [14] therefore carry particular weight in judging readiness for drug discovery use.

*Recommendations*

*For Model Developers*

Model developers should treat scaffold, temporal, and external validation as routine rather than optional additions to random cross-validation. The contrast between benchmark-style evaluation [1, 16], industrial prospective validation [12], and cross-chemical-space graph neural network testing [13] indicates that validation design can be as important as algorithm selection. Developers should also define applicability domains, report uncertainty where possible, and provide sufficient code, data, and preprocessing detail to allow independent replication [14].

**Table 3** defines translational readiness criteria for ML-based ADMET prediction and clarifies why high retrospective accuracy alone is insufficient for drug discovery decision support.

**Table 3.** Translational Readiness Criteria for ML-Based ADMET Prediction Models

| Readiness criterion | Minimum acceptable evidence | Stronger translational evidence | Risk when absent | Relevance to ADMET decision-making |
|---|---|---|---|---|
| Dataset provenance | Clear description of assay source, endpoint definition, compound inclusion, and preprocessing | Public or auditable datasets with documented curation, duplicate handling, and endpoint harmonization | Hidden bias, assay inconsistency, non-reproducible performance | Determines whether model outputs reflect reliable ADMET biology or dataset artifacts. |
| Split strategy | Internal train-test split or k-fold cross-validation reported transparently | Scaffold split, temporal split, external test set, or prospective validation | Inflated performance due to analogue leakage or chemical-space overlap | Critical for estimating how models behave on new compound series. |
| External validation | Independent dataset used when available | Multi-source, temporal, industrial, or laboratory-confirmed external testing | Poor generalizability outside the original training chemistry | Essential for judging whether a model can support real compound prioritization. |

| | | | | |
|---|---|---|---|---|
| Prospective validation | Retrospective validation only, with limitations acknowledged | Predictions tested prospectively against new experimental ADMET results | Deployment without evidence of real-world decision value | Strongest indicator of drug discovery utility. |
| Applicability domain | Basic statement of chemical-space limits | Formal domain estimation, out-of-domain warnings, similarity thresholds, or confidence limits | Overconfident predictions for structurally novel molecules | Prevents unsafe interpretation of predictions outside model experience. |
| Uncertainty reporting | Point estimates or class probabilities reported | Calibrated uncertainty, confidence intervals, ensemble disagreement, or reliability assessment | False precision and unsupported decision confidence | Helps users decide when experimental confirmation should be prioritized. |
| Interpretability | Limited feature importance or endpoint rationale | Mechanistic explanation, molecular substructure attribution, interpretable alerts, or user-facing rationale | Black-box recommendations with limited medicinal chemistry usefulness | Supports medicinal chemistry reasoning and liability investigation. |
| Reproducibility | Metrics and model type described | Code, data splits, preprocessing scripts, hyperparameters, and benchmark comparisons available | Results cannot be independently verified or compared | Enables cumulative progress and fair comparison across ADMET models. |
| Context of use | General prediction purpose stated | Explicit decision role, intended users, acceptable error tolerance, and documentation requirements defined | Misuse as a stand-alone safety or PK decision tool | Distinguishes exploratory screening from high-stakes translational or regulatory use. |

*For Journal Editors*

Journal editors should require ADMET modeling papers to describe dataset provenance, duplicate handling, split strategy, external validation, and applicability-domain assessment in enough detail for readers to judge bias. Papers presenting new deep learning architectures, including graph attention models [8], hybrid graph-fingerprint models [9], multitask deep featurization [29], or deep pharmacokinetic platforms [11], should be evaluated against realistic baselines and chemically meaningful test splits. At minimum, studies should justify why their validation design is appropriate for the claimed drug discovery use case [12, 13].

*For the Community*

The ADMET community should develop curated, living benchmarks with transparent assay definitions, standardized train-test splits, endpoint-specific metadata, and periodic challenge rounds. MoleculeNet demonstrated the value of common molecular machine learning datasets [1], while PharmaBench suggests that ADMET-specific benchmark expansion remains an active need [16]. Such resources should include absorption, distribution, metabolism, excretion, and toxicity endpoints represented in current platforms [3–5, 10], while also documenting applicability-domain boundaries and dataset updates over time [14].

*For Regulators*

Regulatory stakeholders should begin defining expectations for ML-based ADMET evidence, particularly when predictions are used to support candidate nomination, safety assessment, or reduced experimental testing. Existing tools such as SwissADME [2], vNN-ADMET [19], ADMETboost [15], and Interpretable-ADMET [18] illustrate the growing accessibility of model outputs, but accessibility does not equal regulatory reliability. Clear documentation of training data, validation design, uncertainty, interpretability, and intended context of use will be essential before ML-based ADMET predictions can support high-stakes submissions [14, 17].

*Research Gaps*

*Prospective Implementation Studies*

Prospective implementation remains the most important evidence gap in ML-based ADMET prediction. The industrial validation study by Fang, Wang, Grater, Kapadnis, Black, Trapa, and Sciabola [12] shows how prospective assessment can provide stronger evidence than retrospective benchmarking, but comparable studies are uncommon across the broader literature. Future work should report whether predicted ADMET liabilities changed compound prioritization and whether subsequent experiments confirmed or refuted the model-guided decisions [10, 15, 18].

*8Multi-Endpoint and Systems ADMET Models*

Multi-endpoint models are increasingly available, but systems-level ADMET reasoning remains underdeveloped. Platforms such as ADMETlab 2.0 [4], ADMETlab 3.0 [5], ADMET-AI [10], and Deep-PK [11] can generate broad profiles, yet most evidence still evaluates endpoints separately rather than assessing whether combined predictions improve medicinal chemistry decisions. More research is needed on integrated models that jointly represent absorption, distribution, metabolism, excretion, and toxicity trade-offs while preserving interpretability and applicability-domain safeguards [14, 18].

*Transferability across Chemical Space*

Transferability across chemical space remains uncertain, especially for compounds outside conventional drug-like chemistry. Graph neural network benchmarks across different ADME datasets showed that chemical-space shifts can affect generalizability [13], and learned molecular representations do not automatically solve the problem of domain mismatch [7]. Future studies should test models on emerging modalities, macrocycles, PROTAC-like molecules, peptides, and chemically novel screening collections, rather than assuming that models trained on traditional small molecules will transfer reliably [17, 29].

*Implications*

*For Research Practice*

For research practice, the central implication is that validation rigor must become a first-order design choice in ADMET modeling studies. New architectures such as graph attention networks [8], FP-GNN [9], and multitask deep featurization [29] are valuable, but their practical significance depends on validation against realistic chemical novelty. Researchers should therefore report not only model metrics, but also split rationale, endpoint curation, applicability domain, uncertainty, and failure modes [12–14].

*For Drug Discovery*

For drug discovery, current ML-based ADMET models are best interpreted as decision-support tools that can prioritize experimental testing and flag likely liabilities. Web-accessible platforms such as SwissADME [2], admetSAR 2.0 [3], ADMETboost [15], Interpretable-ADMET [18], ADMET-AI [10], and Deep-PK [11] can accelerate early evaluation, but their outputs should not replace in-vitro or in-vivo confirmation. This is especially important for toxicity and metabolism endpoints, where mechanistic complexity, assay context, and chemical-space uncertainty can limit purely computational inference [21, 22, 24, 25].

*For Policy*

For policy, the growth of ML-based ADMET prediction implies a need for clearer reporting and documentation standards. Benchmarking resources [1, 16], industrial validation studies [12], and applicability-domain frameworks [14] provide useful starting points for defining what reliable evidence should include. As ML outputs become more common in internal decision packages and potential regulatory submissions, policy guidance should distinguish exploratory screening predictions from validated models intended to support consequential safety or pharmacokinetic claims [17, 28].

## Conclusion

Machine learning-based ADMET prediction expanded substantially from 2017 to 2024, with absorption and toxicity endpoints among the most actively modeled areas. The field now includes classical machine learning, multitask neural networks, graph neural networks, deep molecular representations, and accessible web-based platforms.

The main barrier to translational impact is not the absence of algorithms, but the inconsistent use of rigorous validation. External testing, temporal splitting, scaffold splitting, prospective validation, uncertainty quantification, and applicability-domain reporting remain too uncommon for many models to be trusted in high-stakes decisions.

Without stronger validation norms and community benchmarks, the growing number of ADMET models risks producing academic performance claims that do not translate into reliable compound-selection tools. A systematic shift toward transparent datasets, realistic split strategies, open reporting, and prospective testing is needed.

The most useful future ADMET models will be those that communicate both predictions and limits. When models are benchmarked honestly, validated externally, and integrated with experimental workflows, they can become powerful decision-support tools for safer and more efficient drug discovery.

**Acknowledgments:** None

**Conflict of interest:** None

**Financial support:** None

**Ethics statement:** None

## References

1. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30.
2. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017;7(1):42717.

3. Yang H, Lou C, Sun L, Li J, Cai Y, Wang Z, et al. admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. Bioinformatics. 2019;35(6):1067-9.
4. Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res. 2021;49(W1):W5-14.
5. Fu L, Shi S, Yi J, Wang N, He Y, Wu Z, et al. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. Nucleic Acids Res. 2024;52(W1):W422-31.
6. Wenzel J, Matter H, Schmidt F. Predictive multitask deep neural network models for ADME-Tox properties: learning from large data sets. J Chem Inf Model. 2019;59(3):1253-68.
7. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88.
8. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem. 2019;63(16):8749-60.
9. Cai H, Zhang H, Zhao D, Wu J, Wang L. FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction. Brief Bioinform. 2022;23(6):bbac408.
10. Swanson K, Walther P, Leitz J, Mukherjee S, Wu JC, Shivnaraine RV, et al. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. Bioinformatics. 2024;40(7):btae416.
11. Myung Y, de Sá AG, Ascher DB. Deep-PK: deep learning for small molecule pharmacokinetic and toxicity prediction. Nucleic Acids Res. 2024;52(W1):W469-75.
12. Fang C, Wang Y, Grater R, Kapadnis S, Black C, Trapa P, et al. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: an industrial perspective. J Chem Inf Model. 2023;63(11):3263-74.
13. Broccatelli F, Trager R, Reutlinger M, Karypis G, Li M. Benchmarking accuracy and generalizability of four graph neural networks using large in vitro ADME datasets from different chemical spaces. Mol Inform. 2022;41(8):2100321.
14. Kar S, Roy K, Leszczynski J. Applicability domain: a step toward confident predictions and decidability for QSAR modeling. In: Roy K, editor. Comput Toxicol Methods Protoc. New York (NY): Springer; 2018. p. 141-69.
15. Tian H, Ketkar R, Tao P. ADMETboost: a web server for accurate ADMET prediction. J Mol Model. 2022;28(12):408.
16. Niu Z, Xiao X, Wu W, Cai Q, Jiang Y, Jin W, et al. PharmaBench: enhancing ADMET benchmarks with large language models. Sci Data. 2024;11(1):985.
17. Bassani D, Parrott NJ, Manevski N, Zhang JD. Another string to your bow: machine learning prediction of the pharmacokinetic properties of small molecules. Expert Opin Drug Discov. 2024;19(6):683-98.
18. Wei Y, Li S, Li Z, Wan Z, Lin J. Interpretable-ADMET: a web service for ADMET prediction and optimization based on deep neural representation. Bioinformatics. 2022;38(10):2863-71.
19. Schyman P, Liu R, Desai V, Wallqvist A. vNN web server for ADMET predictions. Front Pharmacol. 2017;8:889.
20. Ryu JY, Lee MY, Lee JH, Lee BH, Oh KS. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. Bioinformatics. 2020;36(10):3049-55.
21. Wang D, Liu W, Shen Z, Jiang L, Wang J, Li S, et al. Deep learning based drug metabolites prediction. Front Pharmacol. 2020;10:1586.
22. Cai C, Guo P, Zhou Y, Zhou J, Wang Q, Zhang F, et al. Deep learning-based prediction of drug-induced cardiotoxicity. J Chem Inf Model. 2019;59(3):1073-84.
23. Ylipää E, Chavan S, Bånkestad M, Broberg J, Glinghammar B, Norinder U, et al. hERG-toxicity prediction using traditional machine learning and advanced deep learning techniques. Curr Res Toxicol. 2023;5:100121.
24. Chen Z, Jiang Y, Zhang X, Zheng R, Qiu R, Sun Y, et al. The prediction approach of drug-induced liver injury: response to the issues of reproducible science of artificial intelligence in real-world applications. Brief Bioinform. 2022;23(4):bbac196.
25. Litsa EE, Das P, Kavraki LE. Machine learning models in the prediction of drug metabolism: challenges and future perspectives. Expert Opin Drug Metab Toxicol. 2021;17(11):1245-7.
26. Zhu K, Huang M, Wang Y, Gu Y, Li W, Liu G, et al. MetaPredictor: in silico prediction of drug metabolites based on deep language models with prompt engineering. Brief Bioinform. 2024;25(5):bbae374.
27. Wu Z, Lei T, Shen C, Wang Z, Cao D, Hou T. ADMET evaluation in drug discovery. 19. Reliable prediction of human cytochrome P450 inhibition using artificial intelligence approaches. J Chem Inf Model. 2019;59(11):4587-601.
28. Tran TT, Tayara H, Chong KT. Artificial intelligence in drug metabolism and excretion prediction: recent advances, challenges, and future perspectives. Pharmaceutics. 2023;15(4):1260.
29. Feinberg EN, Joshi E, Pande VS, Cheng AC. Improvement in ADMET prediction with multitask deep featurization. J Med Chem. 2020;63(16):8835-4.