Pharmacophore

ISSN-2229-5402

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

SELF-SUPERVISED MOLECULAR MODELS FOR P-GLYCOPROTEIN SUBSTRATE PREDICTION USING TRANSPORTER ASSAY DATA

Thomas Andersen[1]*, Lars Nielsen[1], Mette Sørensen[2]

1. ***Department of Pharmaceutical Informatics and AI, Faculty of Pharmacy, University of Copenhagen, Copenhagen, Denmark.***
2. ***Department of Computational Pharmacology, Faculty of Engineering, Technical University of Denmark, Lyngby, Denmark.***

ARTICLE INFO

Received:
23 November 2025
Received in revised form:
10 February 2026
Accepted:
12 February 2026
Available online:
28 February 2026

Keywords: Self-supervised learning, P-glycoprotein, ABCB1, Transporter assays, Molecular graphs, ADME

ABSTRACT

P-glycoprotein efflux can strongly constrain oral absorption, brain penetration, and intracellular drug exposure. Computational substrate prediction is therefore an important early filter for molecules likely to face transporter-mediated disposition liabilities. Most transporter models rely on limited labeled assay data and are often trained directly on endpoint-specific measurements. This ignores the broader chemical information contained in large collections of unlabeled molecular structures. This MDL article proposes a self-supervised molecular model for P-glycoprotein substrate prediction. The model pre-trains on large unlabeled chemical databases and is then adapted to a limited set of validated transporter assay labels. A molecular encoder would be pre-trained using contrastive and masked-structure objectives over graph or SMILES representations. The pre-trained encoder would then be coupled to a lightweight classifier for binary substrate prediction using curated P-glycoprotein assay labels.
Conceptually, the self-supervised model would be expected to offer better data efficiency than a model trained only from limited labeled transporter data. Attribution methods could also highlight molecular features associated with P-glycoprotein recognition. Self-supervised molecular learning could make transporter prediction more accessible when labeled assay data are scarce. This approach may support earlier design of molecules with more favorable absorption and distribution profiles.

To Cite This Article: Andersen T, Nielsen L, Sørensen M. Self-Supervised Molecular Models for P-Glycoprotein Substrate Prediction Using Transporter Assay Data. Pharmacophore. 2026;17(1):91-100. https://doi.org/10.51847/SLmtXHWnZI

Introduction

P-glycoprotein is a central efflux transporter in drug disposition because it can reduce intracellular drug exposure and restrict tissue penetration, particularly across epithelial and barrier tissues. Predicting whether a molecule is a P-glycoprotein substrate is therefore an important computational screening task before extensive experimental investment. Recent transporter modeling studies have used assay-derived labels and machine learning to classify substrate or interaction potential [1], while in silico approaches focused on brain penetration illustrate how efflux liability can shape distribution-relevant decisions [2]. For an MDL framework, P-glycoprotein substrate prediction is best treated not as a purely descriptive QSAR problem, but as a representation-learning problem grounded in transporter biology.

The key limitation is that experimentally reliable transporter labels remain sparse relative to the size of chemical space. Bidirectional transport, ATPase activation, and uptake-based assays can provide informative substrate annotations, but their outputs are sensitive to assay context and curation decisions [3]. Public transporter resources have begun to consolidate interaction data across compounds and transporters [4], and broader chemical-transporter profiling platforms further demonstrate the value of integrating scattered transporter evidence [5]. However, even curated collections remain small compared with the unlabeled molecules available for chemical representation learning.

Self-supervised molecular learning addresses this mismatch by learning transferable chemical features before endpoint-specific labels are introduced. Molecular benchmark resources helped standardize evaluation of learned representations across property prediction tasks [6], while unsupervised approaches such as Mol2vec showed that chemical structure can be encoded

Corresponding Author: Thomas Andersen; Department of Pharmaceutical Informatics and AI, Faculty of Pharmacy, University of Copenhagen, Copenhagen, Denmark. E-mail: thomas.andersen@gmail.com.

from unlabeled molecular fragments with chemically meaningful behavior [7]. Graph neural networks and learned molecular embeddings then extended this idea by allowing models to infer task-relevant patterns directly from molecular topology [8, 9]. These developments make P-glycoprotein substrate prediction a natural downstream target for pre-trained molecular encoders.

The proposed model follows a transfer-learning logic in which an encoder is first exposed to broad unlabeled chemical structure and then fine-tuned on limited transporter assay labels. This design aligns with modern molecular pre-training strategies that learn expressive global molecular representations for drug discovery [10] and with SMILES-based language models that adapt transformer representations to molecular property prediction [11]. For P-glycoprotein specifically, the model should combine the broad chemical priors learned from unlabeled databases with endpoint-specific evidence from validated substrate assays [12]. The result is a conceptual MDL architecture intended to predict substrate likelihood without claiming experimental validation or numerical performance in this manuscript.

Background

P-Glycoprotein Biology and Its Role in ADME

P-glycoprotein, encoded by ABCB1, functions as an ATP-dependent efflux transporter that can lower intracellular concentrations of structurally diverse xenobiotics. Its biological role is especially relevant to absorption, distribution into protected tissues, and resistance phenotypes, making it a recurring liability in drug discovery. Computational studies that target P-glycoprotein substrates and modulators underscore how transporter recognition can be framed as a molecular classification problem [13]. Structure-aware and docking-supported modeling has also been used to relate ligand properties to potential interaction with the transporter binding region [14].

In-Vitro Transporter Assays for Substrate Identification

P-glycoprotein substrate labels are commonly inferred from in-vitro systems such as polarized cell monolayers, efflux ratio measurements, ATPase stimulation, and fluorescent probe displacement or uptake assays. Simplified experimental workflows and associated in silico models show that substrate assignment depends on how transport potential is operationalized from assay readouts [3]. Studies using brain capillary endothelial models further show that assay context can affect how efflux potential is translated into disposition-relevant interpretation [2]. These considerations make careful label curation essential for any fine-tuned transporter model.

Traditional Machine Learning and QSAR for P-gp Prediction

Earlier P-glycoprotein models often relied on curated descriptors, fingerprints, or engineered molecular features coupled with classical classifiers. Decision tree approaches made explicit chemical rules accessible for P-glycoprotein substrate and inhibitor prediction [13], while heterogeneous classifier fusion showed that combining model families can improve robustness when transporter data are noisy [1]. Support vector regression and related schemes also illustrate how physicochemical descriptors can be mapped to efflux-related endpoints [15]. Although useful, these approaches may be constrained by handcrafted feature sets and by limited labeled data.

Self-Supervised and Contrastive Learning on Molecules

Self-supervised molecular learning trains encoders to recover structural information from molecules without relying on endpoint labels. Contrastive objectives encourage related molecular views to have consistent embeddings, while masked atom or token modeling asks the model to infer hidden local chemical context from the remaining structure. Recent self-supervised molecular frameworks have used these principles to learn expressive global representations [10], and masked graph transformer autoencoders extend this idea by reconstructing molecular information from corrupted graph inputs [16]. Multimodal and contrastive molecular frameworks further suggest that complementary graph, sequence, and conformational views can strengthen transferable chemical representations [17, 18].

Transfer and Few-Shot Learning in ADME

ADME endpoints often have fewer high-quality labels than standard molecular property benchmarks, which makes transfer learning attractive. General-purpose ADMET platforms demonstrate how multiple property predictors can be organized for early drug discovery screening [19], while Therapeutics Data Commons provides a structured framework for evaluating drug discovery models across clinically relevant tasks. Pre-trained molecular representations are especially useful when a downstream endpoint, such as P-glycoprotein substrate status, has limited curated labels but strong dependence on general chemical features. In this setting, fine-tuning should adapt a broad chemical encoder to transporter-specific decision boundaries rather than learning all features from scratch.

Model Development Overview

High-Level Training Pipeline

The proposed pipeline begins with a molecular encoder trained on large unlabeled chemical corpora and then adapts that encoder to binary P-glycoprotein substrate prediction. This mirrors the broader movement from isolated endpoint models toward pre-trained molecular systems that support downstream property prediction [20]. After pre-training, the encoder would

be connected to a compact classification head and fine-tuned using curated transporter labels from assays or transporter knowledge resources [4]. The architecture is conceptual and is intended to define how such a model should be built and evaluated, not to report completed experiments.

Figure 1 presents the proposed self-supervised molecular learning architecture in which broad unlabeled chemical structures are used to pre-train a transferable encoder that is subsequently fine-tuned on curated P-glycoprotein transporter assay labels for calibrated substrate prediction and interpretable medicinal chemistry feedback.

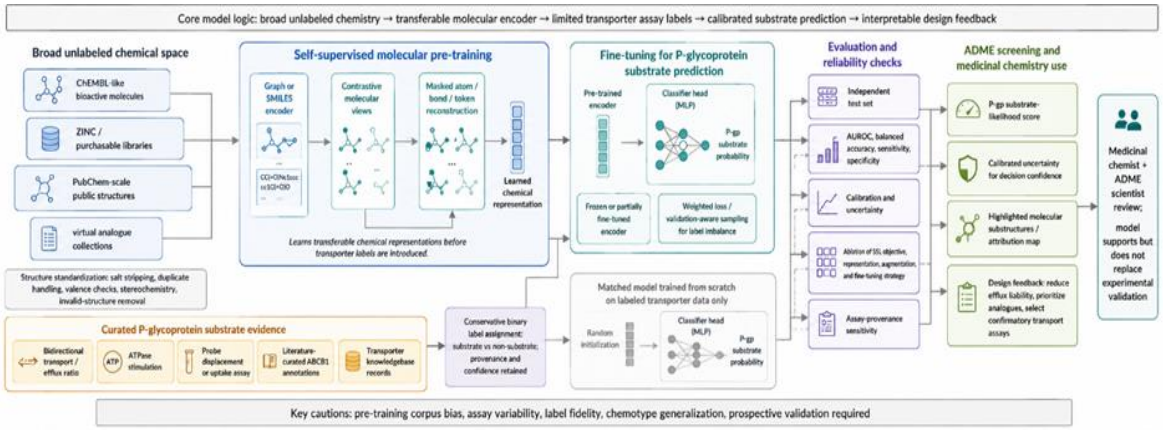

Figure 1. Self-Supervised Molecular Learning Architecture for P-Glycoprotein Substrate Prediction from Unlabeled Chemistry and Transporter Assay Labels

Core Input Representations and Tasks

The model can accept molecular graphs, SMILES strings, or paired representations depending on the encoder family selected. Graph-based encoders capture atom-bond topology directly, while SMILES-based language models use tokenized sequences to learn chemical grammar and context [11]. Transformer studies comparing molecular tokenization strategies indicate that representation choices can affect how chemical syntax and semantics are learned [21]. For the downstream task, each molecule would be paired with a binary substrate or non-substrate label derived from curated P-glycoprotein assay interpretation.

Design Principles

The design should prioritize data efficiency, representation transferability, and interpretability because transporter endpoints are label-limited and biologically nuanced. Molecular graph attention architectures demonstrate how learned representations can emphasize chemically relevant neighborhoods [22], while interpretable P-glycoprotein graph models show that prediction and explanation can be developed together for transporter classification [23]. The model should remain modality-agnostic enough to incorporate graphs, SMILES, fingerprints, or conformer-derived features when appropriate. Its output should be a calibrated substrate-likelihood estimate accompanied by molecular explanations suitable for medicinal chemistry review.

Table 1 maps the proposed model from unlabeled molecular representation learning to transporter-specific fine-tuning, clarifying how each architectural layer contributes to data-efficient P-glycoprotein substrate prediction.

Table 1. Self-Supervised Molecular Representation-to-Transporter Prediction Design Map

Model design layer	Scientific purpose in this manuscript	Candidate implementation choices	Why this layer matters for P-glycoprotein substrate prediction	Main risk if poorly specified
Unlabeled chemical pre-training corpus	Provides broad chemical exposure before transporter labels are introduced	ChEMBL-like bioactive molecules; ZINC or ZINC20-like purchasable and virtual compounds; PubChem-scale public structures; internal virtual analogue libraries	Allows the encoder to learn transferable structural regularities from chemical space that is much larger than the labeled P-glycoprotein assay set	The encoder may learn biased chemical priors if the corpus overrepresents common drug-like structures and underrepresents macrocycles, natural products, peptides, or atypical chemotypes
Molecular standardization layer	Ensures that molecular inputs represent valid and comparable chemical entities	Salt stripping; tautomer handling; duplicate removal; valence checks; stereochemical annotation; invalid-structure exclusion	Reduces noise before graph or SMILES representations are learned and prevents technical artifacts from becoming predictive features	Inconsistent standardization can cause the same molecule to appear as multiple conflicting structures or can remove substrate-relevant stereochemical information
Input representation	Defines how chemical structure	Molecular graph; SMILES sequence; paired graph–	P-glycoprotein recognition may depend on both local	A narrow representation may miss topology,

	is exposed to the encoder	SMILES representation; optional conformer-aware descriptors	substructures and distributed physicochemical patterns, so the representation must preserve chemically meaningful context	stereochemistry, or sequence context relevant to transporter recognition
Self-supervised objective	Teaches the encoder to learn molecular structure without endpoint labels	Contrastive learning between augmented molecular views; masked atom, bond, or token reconstruction; hybrid contrastive–generative pre-training	Builds a chemical prior that can later be adapted to scarce transporter labels rather than learning all structure–activity patterns from scratch	Poorly chosen augmentations may create unrealistic molecular views or teach invariances that erase substrate-relevant motifs
Molecular encoder backbone	Converts each molecule into a reusable latent representation	Graph neural network; graph attention network; graph transformer; SMILES transformer; hybrid encoder	Produces the embedding that supports downstream binary substrate classification and atom-level explanation	Excessive model complexity may overfit during fine-tuning, while insufficient capacity may fail to capture distributed substrate-recognition patterns
Labeled transporter assay layer	Supplies endpoint-specific evidence for adapting the pre-trained encoder	Bidirectional transport assays; efflux ratios; ATPase stimulation; probe displacement or uptake assays; curated ABCB1 substrate annotations	Grounds the downstream classifier in experimentally interpreted P-glycoprotein substrate evidence	Mixed assay types may produce inconsistent labels if provenance, thresholding, and confidence are not retained
Fine-tuning strategy	Adapts the broad chemical encoder to the specific transporter endpoint	Frozen encoder followed by classifier training; partial fine-tuning; weighted loss; validation-aware sampling; few-shot adaptation	Supports data-efficient learning when high-confidence substrate labels are limited	Full fine-tuning on small data may erase useful pre-training knowledge or amplify assay-specific bias
Baseline comparison	Tests whether self-supervision provides added value over supervised-only learning	Same architecture trained from random initialization; ECFP-based classifier; descriptor-based QSAR; classical machine learning baseline	Distinguishes genuine transfer benefit from model-capacity or data-split effects	Without matched baselines, performance claims may incorrectly attribute gains to pre-training
Output representation	Converts model inference into usable ADME screening information	Substrate-likelihood score; calibrated uncertainty; binary classification threshold; ranked compound list	Enables medicinal chemists and ADME scientists to triage analogues and prioritize confirmatory transporter assays	A poorly calibrated score may be mistaken for biological certainty rather than model-estimated likelihood
Interpretation layer	Translates predictions into chemically reviewable explanations	SHAP; integrated gradients; attention analysis; fragment masking; substructure attribution	Supports identification of molecular regions associated with predicted P-glycoprotein recognition or efflux liability	Explanations may appear plausible but fail to correspond to chemically meaningful or experimentally testable motifs

Data Sources and Feature Engineering

Unlabeled Pre-Training Corpus

The unlabeled pre-training corpus would be assembled from broad chemical databases after standardization, salt stripping, valence checks, duplicate handling, and removal of chemically invalid structures. ChEMBL provides bioactive molecules with drug discovery relevance [24], ZINC supports ligand discovery through purchasable and virtual compound collections [25], and ZINC20 expands this role for ultralarge-scale ligand exploration [26]. PubChem contributes a broad public chemical repository that can further diversify pre-training chemistry [27]. The purpose of this corpus is not to provide transporter labels, but to expose the encoder to general chemical structure before downstream adaptation.

Labeled P-gp Substrate Dataset

The labeled P-glycoprotein dataset would be assembled from literature-curated transporter assays and public transporter interaction resources using conservative rules for binary substrate assignment. Assay-derived models for P-glycoprotein transport potential show how experimental outputs can be converted into computational labels when substrate evidence is sufficiently clear [3]. TICBase provides an integrated source for compound-transporter interaction data [4], while broader

transporter profilers such as MONSTROUS support cross-transporter interaction organization [5]. Label curation should preserve assay provenance so that fine-tuning can distinguish high-confidence substrate evidence from ambiguous interaction records.

Molecular Graph Construction and Augmentation

Molecular graph construction would convert standardized structures into atoms, bonds, formal charges, aromaticity indicators, stereochemical descriptors, and other chemically valid features. Learned molecular representation studies show that graph-based inputs can support property prediction when atom and bond context are encoded systematically [9]. Augmentations should create alternative views through chemically cautious perturbations such as masking atoms, hiding substructures, or modifying graph neighborhoods without generating invalid molecules. Such augmentation choices are central to contrastive molecular learning because the model should become invariant to superficial view changes while remaining sensitive to substrate-relevant chemistry [10].

Self-Supervised Pre-Training Architecture

Graph Encoder Backbone

A graph neural network backbone such as a graph isomorphism network or graph attention network would transform each molecular graph into a fixed-length embedding. Attention-based graph methods have shown that molecular neighborhoods can be weighted differently during property prediction [22], and potential-based neural architectures demonstrate the value of learned interaction-aware molecular features [8]. For P-glycoprotein substrate prediction, the backbone should capture both local motifs and broader molecular organization because transporter recognition may depend on distributed physicochemical patterns. The encoder would therefore be designed to preserve atom-level information for attribution while producing molecule-level embeddings for classification.

Contrastive and Generative Pre-Training Objectives

The pre-training objective would combine contrastive learning between augmented molecular views with generative reconstruction of masked atoms, bonds, or tokens. Self-supervised molecular pre-training studies indicate that reconstructive and contrastive tasks can help models learn transferable representations before downstream labels are introduced [28]. Masked graph transformer autoencoding further supports the idea that corrupted molecular graphs can train encoders to recover chemically meaningful structure [16]. SMILES and conformer-aware approaches provide complementary evidence that both sequence context and molecular geometry can enrich property-oriented molecular embeddings [29, 30].

Pre-Training Scale and Infrastructure

Pre-training would use scalable molecular learning infrastructure to process broad unlabeled corpora and produce a reusable encoder for downstream ADME tasks. The conceptual emphasis is on breadth and diversity of chemical exposure rather than on reporting specific compute, epochs, or dataset counts. Reviews of molecular transformers show that architecture, tokenization, and pre-training objective choices can influence transfer behavior across drug discovery tasks [20], while multimodal self-supervised frameworks suggest that combining complementary molecular views may improve representation robustness [18]. The resulting encoder would serve as a general-purpose chemical prior that can be adapted to P-glycoprotein substrate prediction using limited labeled assay data.

Fine-Tuning for P-gp Substrate Prediction

Transfer Learning Setup

For P-glycoprotein substrate prediction, the pre-trained encoder would be transferred to a supervised binary classification task by attaching a lightweight multilayer classifier to the molecular embedding. The encoder could be frozen initially to preserve general chemical knowledge, then partially fine-tuned so that higher layers adapt to transporter-specific patterns in curated substrate labels. A multimodal contrastive P-glycoprotein framework illustrates how pre-trained molecular information can be specialized for substrate and inhibitor prediction tasks [31]. The supervised objective should therefore refine a broad chemical representation into a substrate-likelihood function without requiring the model to learn molecular chemistry entirely from limited transporter labels.

Handling Class Imbalance and Limited Data

P-glycoprotein datasets are likely to contain uneven representation of substrates and non-substrates because experimental testing is often biased toward compounds already suspected of transporter interaction. Curated machine learning workflows for ABC transporter efflux and inhibition emphasize the importance of data curation, interaction labeling, and endpoint-specific modeling when transporter annotations are heterogeneous [32]. Weighted loss functions, cautious augmentation, and validation-aware sampling would be appropriate conceptual strategies for reducing bias during fine-tuning. Meta-learning or few-shot adaptation could also be considered when the labeled set is too narrow to support extensive endpoint-specific training.

Comparing Models Trained From Scratch

The central comparison should be between the SSL-pre-trained encoder and the same architecture initialized randomly and trained only on labeled P-glycoprotein data. Such a comparison would test whether the representation learned from unlabeled chemistry provides a meaningful inductive bias for transporter prediction. Benchmarking principles from molecular machine learning indicate that fair comparisons require consistent splits, matched model capacity, and careful baseline selection [6]. Generative molecular benchmarks also show that evaluation should separate representation quality from downstream optimization artifacts, which is relevant when judging whether pre-training genuinely improves substrate classification [33].

Model Interpretability and Substrate Pharmacophore Identification

Substrate-Specific Feature Attribution

Model interpretation would focus on identifying molecular regions that contribute most strongly to a positive P-glycoprotein substrate prediction. Attribution methods such as SHAP, attention analysis, integrated gradients, or fragment masking could highlight basic centers, lipophilic groups, hydrogen-bonding patterns, or distributed structural motifs that the model associates with transporter recognition. Interpretable graph neural network protocols for P-glycoprotein prediction provide a relevant template for linking model decisions to chemically meaningful substructures [23]. Docking-supported machine learning studies also suggest that interpretation can be strengthened when ligand-based attributions are compared with plausible transporter interaction regions [14].

From Explanation to Chemical Rule

Repeated attribution of similar fragments across chemically diverse predicted substrates could support extraction of a learned pharmacophore. This pharmacophore should be treated as a model-derived design hypothesis rather than as a definitive mechanistic rule, because transporter recognition may involve flexible binding and multiple interaction modes. Decision tree approaches such as PgpRules demonstrate how explicit rule systems can make P-glycoprotein predictions more transparent for users [13]. A self-supervised graph model could extend this idea by learning rules from distributed embeddings while still translating them into medicinal chemistry guidance.

Integration into ADME Screening Workflows

Virtual Screening of Compound Libraries

The fine-tuned model could be deployed as an early ADME screening component that assigns substrate-likelihood scores to proposed analogues before in-vitro transporter testing. This would allow medicinal chemistry teams to prioritize compounds with lower predicted efflux liability or to flag molecules requiring confirmatory transport assays. Integrated ADMET prediction platforms show how computational property models can support early triage across absorption, distribution, metabolism, excretion, and toxicity endpoints [19]. A P-glycoprotein-specific SSL model would fit into this workflow as a focused transporter-risk module rather than as a replacement for experimental evaluation.

Prospective Experimental Validation

Prospective validation should involve selecting compounds predicted as likely substrates and likely non-substrates, then testing them in standardized bidirectional transport or related assays. Simplified P-glycoprotein substrate assay workflows provide a practical foundation for linking model predictions to experimental transport potential [3]. Brain penetration-oriented efflux models also show how prospective transporter interpretation can be connected to tissue-distribution questions [2]. In a discovery setting, validation should be iterative, with newly generated assay labels used to refine the fine-tuned classifier and assess whether the model generalizes beyond the chemistry used during training.

Evaluation Strategy

Predictive Performance

Predictive evaluation should use an independent test set and report threshold-free and threshold-dependent classification behavior, including AUROC, balanced accuracy, sensitivity, and specificity, without relying on any single metric. ECFP-based classifiers, descriptor-based QSAR models, and graph models trained from scratch would provide appropriate baselines for determining whether pre-training improves transporter prediction. Prior P-glycoprotein substrate and inhibitor models demonstrate the range of classical and deep learning baselines that should be considered in such comparisons [1, 12, 15]. The evaluation should also assess calibration, because a substrate-likelihood score is most useful when it reflects meaningful uncertainty for ADME decision-making.

Ablation Studies

Ablation studies should isolate the contribution of contrastive learning, masked reconstruction, input representation, and fine-tuning strategy. Triple generative self-supervised learning illustrates how multiple pre-training tasks can be combined for molecular property prediction, making objective-level ablation especially important [34]. SMILES tokenization comparisons also indicate that sequence representation choices can influence chemical language modeling behavior and should be tested separately from graph-based design choices [21]. These analyses would clarify whether performance gains arise from self-supervision itself, from specific molecular augmentations, or from broader chemical coverage in pre-training.

Interpretability Benchmark

Interpretability evaluation should ask whether highlighted substructures align with established transporter pharmacophores and with expert medicinal chemistry expectations. Attention-based molecular property models show that graph neural networks can assign different importance to molecular neighborhoods [22], while P-glycoprotein-specific interpretable graph protocols provide a more endpoint-focused benchmark for explanation quality [23]. Expert review could compare model-highlighted fragments with substrate-associated motifs across congeneric series, while prospective SAR follow-up could test whether modifying highlighted groups changes predicted or observed transporter behavior. This would make interpretability a practical design tool rather than a purely visual explanation.

Table 2 provides an evaluation and deployment-readiness framework that separates predictive performance, calibration, interpretability, assay robustness, chemical-space generalization, and prospective validation requirements for the proposed self-supervised P-glycoprotein model.

Table 2. Evaluation, Interpretability, and Deployment Readiness Framework for a Self-Supervised P-Glycoprotein Substrate Model

Evaluation domain	Core question	Recommended analysis	Decision-relevant interpretation	Failure mode detected	Practical implication for ADME deployment
Predictive discrimination	Does the model separate likely substrates from likely non-substrates?	AUROC, balanced accuracy, sensitivity, specificity, precision–recall analysis, threshold-specific confusion matrices	Determines whether the classifier can rank or classify compounds for early transporter-risk triage	Apparent high accuracy driven by class imbalance or overrepresentation of common chemotypes	Use model outputs only if discrimination remains stable across independent test data and relevant chemical subgroups
Calibration and uncertainty	Does the substrate-likelihood score behave like a meaningful confidence estimate?	Calibration curve, expected calibration error, Brier score, uncertainty stratification	Indicates whether a predicted probability can guide assay prioritization and decision confidence	Overconfident predictions for molecules outside the training distribution	Flag uncertain predictions for confirmatory transport assays rather than using them for exclusion decisions
Transfer-learning value	Does self-supervised pre-training improve performance beyond supervised-only learning?	Matched comparison against randomly initialized architecture and classical QSAR baselines using identical splits	Tests whether broad unlabeled chemical learning adds useful inductive bias	Performance gain caused by architecture size, data leakage, or favorable split rather than self-supervision	Support use of SSL only if gains persist under fair baseline comparisons
Ablation of pre-training objectives	Which part of self-supervision contributes to downstream prediction?	Remove contrastive objective; remove masked reconstruction; compare graph-only, SMILES-only, and hybrid encoders; vary augmentation strategy	Identifies whether representation gains arise from specific learning objectives or input modalities	Complex pre-training pipeline adds little beyond simpler molecular fingerprints or supervised graph models	Simplify the architecture if performance is not sensitive to advanced SSL components
Assay-provenance sensitivity	Are predictions robust across different substrate-label sources?	Stratified evaluation by assay type, laboratory source, confidence tier, concentration range, and endpoint definition	Determines whether the model learns substrate biology or assay-specific artifacts	Model performs well only on one assay format or curation rule	Retain provenance metadata and avoid merging heterogeneous labels without sensitivity analysis
Chemical-space generalization	Does the model generalize beyond familiar drug-like molecules?	Scaffold split; temporal split; chemotype holdout; distance-to-pre-training-distribution analysis	Shows whether the model can support prospective analogue design rather than memorizing known scaffolds	Strong performance on random splits but weak performance on scaffold or chemotype holdouts	Apply stronger human review for predictions on macrocycles, natural products, peptides, or atypical chemotypes
Interpretability quality	Do highlighted atoms, fragments, or motifs align with plausible transporter-recognition hypotheses?	Expert medicinal chemistry review; fragment masking validation; comparison with known substrate-associated motifs; congeneric-series inspection	Determines whether attributions can support design feedback rather than simply decorate predictions	Attribution maps are unstable, chemically incoherent, or inconsistent across similar compounds	Use explanations as hypotheses for analogue design only when they are stable and chemically plausible
Prospective experimental validity	Do predictions correspond to future transporter assay outcomes?	Select predicted substrates and non-substrates for bidirectional transport or related assays; compare	Provides the strongest evidence that the model is useful for	Retrospective performance does not translate into prospective assay agreement	Deploy as a screening aid only after prospective validation in

		observed assay labels with predicted likelihood	drug discovery decisions		the intended chemical series
Workflow integration	Can the model support practical ADME decision-making without replacing experiments?	Define decision thresholds, uncertainty flags, assay-trigger rules, and review responsibilities	Connects model output to compound triage, analogue prioritization, and confirmatory assay planning	Users overinterpret the model as a definitive substrate assay substitute	Position the system as a decision-support layer for medicinal chemists and ADME scientists
Monitoring and updating	Does model performance remain reliable as new chemistry and assay data accumulate?	Periodic recalibration; drift detection; active learning from new assay labels; versioned model updates	Maintains reliability as compound libraries and assay protocols evolve	Performance decay due to new chemotypes or changed assay practices	Establish a lifecycle plan for model retraining, documentation, and governance

Limitations

Pre-Training Dataset Bias

The unlabeled chemical corpus may be biased toward drug-like, synthetic, and commercially accessible compounds, which could limit generalization to natural products, macrocycles, peptides, or atypical chemotypes. ChEMBL, ZINC, ZINC20, and PubChem offer broad chemical coverage, but each reflects the collection practices, registration history, and intended use cases of its source database [24-27]. A pre-trained encoder may therefore learn chemical priors that are valuable for common medicinal chemistry space while underrepresenting structures outside that space. Bias analysis should examine whether prediction confidence changes for molecules that differ substantially from the pre-training distribution.

Assay Variability and Label Fidelity

P-glycoprotein substrate labels can vary across cell systems, experimental protocols, concentration ranges, and interpretation thresholds. Assay-derived transport potential models show that even simplified substrate classification depends on how experimental outputs are curated and converted into labels [3]. Transporter knowledgebases and interaction profilers help aggregate evidence, but aggregation can also mix assay types that measure related yet nonidentical biology [4, 5]. The model may therefore learn assay-specific patterns unless label provenance, confidence, and endpoint definitions are handled carefully during training and evaluation.

Conclusion

A self-supervised molecular model for P-glycoprotein substrate prediction would address a central imbalance in transporter modeling: the abundance of unlabeled molecular structures and the scarcity of high-confidence transporter assay labels. By pre-training a molecular encoder on broad chemical structure and then adapting it to curated P-glycoprotein labels, the model could learn substrate-relevant representations more efficiently than a purely supervised model.

The main strength of this approach is its compatibility with label-limited ADME prediction. It can combine graph or SMILES-based chemical learning with endpoint-specific fine-tuning, while also producing attribution maps that may help medicinal chemists understand why a compound is predicted to be a substrate. This makes the framework useful not only as a classifier, but also as a hypothesis-generating tool for reducing efflux liability.

Important challenges remain before such a model could be considered reliable in practical discovery settings. Pre-training data may overrepresent familiar drug-like chemistry, transporter labels may contain assay-dependent noise, and prospective validation would still be required to determine whether model-guided decisions improve compound selection. These limitations argue for cautious deployment alongside experimental transporter assays rather than replacement of them.

Future work should support public sharing of molecular encoders tailored to ADME tasks, including transporter-specific adaptation protocols and transparent curation standards. Shared pre-trained models could lower the barrier to high-quality transporter prediction for groups with limited labeled data. With careful validation and interpretability, self-supervised learning could become a practical foundation for more data-efficient drug disposition modeling.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Shaikh N, Sharma M, Garg P. Selective fusion of heterogeneous classifiers for predicting substrates of membrane transporters. J Chem Inf Model. 2017;57(3):594-607.
2. Watanabe R, Esaki T, Ohashi R, Kuroda M, Kawashima H, Komura H, et al. Development of an in silico prediction model for P-glycoprotein efflux potential in brain capillary endothelial cells toward the prediction of brain penetration. J Med Chem. 2021;64(5):2725-38.
3. Ohashi R, Watanabe R, Esaki T, Taniguchi T, Torimoto-Katori N, Watanabe T, et al. Development of simplified in vitro P-glycoprotein substrate assay and in silico prediction models to evaluate transport potential of P-glycoprotein. Mol Pharm. 2019;16(5):1851-63.
4. Michel ME, Wen CC, Yee SW, Giacomini KM, Hamdoun A, Nicklisch SC. TICBase: integrated resource for data on drug and environmental chemical interactions with mammalian drug transporters. Clin Pharmacol Ther. 2023;114(6):1293-303.
5. AbdulHameed MD, Dey S, Xu Z, Clancy B, Desai V, Wallqvist A. MONSTROUS: a web-based chemical-transporter interaction profiler. Front Pharmacol. 2025;16:1498945.
6. 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.
7. Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. J Chem Inf Model. 2018;58(1):27-35.
8. Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, et al. PotentialNet for molecular property prediction. ACS Cent Sci. 2018;4(11):1520-30.
9. 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.
10. Li P, Wang J, Qiao Y, Chen H, Yu Y, Yao X, et al. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. Brief Bioinform. 2021;22(6):bbab109.
11. Li J, Jiang X. Mol-BERT: an effective molecular representation with BERT for molecular property prediction. Wirel Commun Mob Comput. 2021;2021(1):7181815.
12. Esposito C, Wang S, Lange UE, Oellien F, Riniker S. Combining machine learning and molecular dynamics to predict P-glycoprotein substrates. J Chem Inf Model. 2020;60(10):4730-49.
13. Wang PH, Tu YS, Tseng YJ. PgpRules: a decision tree based prediction server for P-glycoprotein substrates and inhibitors. Bioinformatics. 2019;35(20):4193-5.
14. Kadioglu O, Efferth T. A machine learning-based prediction platform for P-glycoprotein modulators and its validation by molecular docking. Cells. 2019;8(10):1286.
15. Chen C, Lee MH, Weng CF, Leong MK. Theoretical prediction of the complex P-glycoprotein substrate efflux based on the novel hierarchical support vector regression scheme. Molecules. 2018;23(7):1820.
16. Wang Z, Feng Z, Li Y, Li B, Wang Y, Sha C, et al. BatmanNet: bi-branch masked graph transformer autoencoder for molecular representation. Brief Bioinform. 2024;25(1):bbad400.
17. Nguyen LD, Nguyen QH, Trinh QH, Nguyen BP. From SMILES to enhanced molecular property prediction: a unified multimodal framework with predicted 3D conformers and contrastive learning techniques. J Chem Inf Model. 2024;64(24):9173-95.
18. Shen A, Yuan M, Ma Y, Du J, Wang M. Complementary multi-modality molecular self-supervised learning via non-overlapping masking for property prediction. Brief Bioinform. 2024;25(4):bbae256.
19. 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.
20. Sultan A, Sieg J, Mathea M, Volkamer A. Transformers for molecular property prediction: Lessons learned from the past five years. J Chem Inf Model. 2024;64(16):6259-80.
21. Leon M, Perezhohin Y, Peres F, Popovič A, Castelli M. Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling. Sci Rep. 2024;14(1):25016.
22. 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.
23. Hsu KC, Wang PH, Su BH, Tseng YJ. A robust and interpretable graph neural network-based protocol for predicting p-glycoprotein substrates. Brief Bioinform. 2025;26(4):bbaf392.
24. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945-54.
25. Tingle BI, Tang KG, Castanon M, Gutierrez JJ, Khurelbaatar M, Dandarchuluun C, et al. ZINC-22─ A free multi-billion-scale database of tangible compounds for ligand discovery. J Chem Inf Model. 2023;63(4):1166-76.
26. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20—a free ultralarge-scale chemical database for ligand discovery. J Chem Inf Model. 2020;60(12):6065-73.
27. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373-80.

28. Wu Z, Cai X, Zhang C, Qiao H, Wu Y, Zhang Y, et al. Self-supervised molecular pretraining strategy for low-resource reaction prediction scenarios. J Chem Inf Model. 2022;62(19):4579-90.
29. Liu Y, Zhang R, Li T, Jiang J, Ma J, Wang P. MolRoPE-BERT: An enhanced molecular representation with Rotary Position Embedding for molecular property prediction. J Mol Graph Model. 2023;118:108344.
30. Qiao J, Jin J, Wang D, Teng S, Zhang J, Yang X, et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. Nat Commun. 2025;16(1):4382.
31. Zhang Y, Wu J, Kang Y, Hou T. A multimodal contrastive learning framework for predicting P-glycoprotein substrates and inhibitors. J Pharm Anal. 2025:101313.
32. Daood NJ, Carey SR, Chung E, Wang T, Kreutz A, Girireddy M, et al. Machine Learning Modeling for ABC Transporter Efflux and Inhibition: Data Curation, Model Development, and New Compound Interaction Predictions. Mol Pharm. 2025;22(11):7022-35.
33. Brown N, Fiscato M, Segler MH, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096-108.
34. Xu L, Xia L, Pan S, Li Z. Triple generative self-supervised learning method for molecular property prediction. Int J Mol Sci. 2024;25(7):3794.