Pharmacophore

ISSN-2229-5402

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

MACHINE LEARNING FOR ORAL BIOAVAILABILITY PREDICTION USING MOLECULAR, PERMEABILITY, METABOLISM, AND FORMULATION FEATURES

Wei Chen[1]*, Li Zhang[1]

1. *Department of AI-Based Pharmaceutical Engineering, Faculty of Pharmacy, School of Pharmaceutical Sciences, Peking University, Beijing, China.*

ARTICLE INFO

Received:
14 September 2024
Received in revised form:
20 November 2024
Accepted:
29 November 2024
Available online:
28 December 2024

Keywords: Oral bioavailability, Machine learning, ADME, Permeability, Intrinsic clearance, Formulation

ABSTRACT

Oral bioavailability is a key determinant of whether a drug candidate can be developed as a practical oral medicine. It reflects the combined influence of molecular structure, intestinal permeability, metabolic extraction, and formulation-dependent release or solubilization. Many prediction approaches rely on simplified molecular rules or isolated in vitro measurements. Such approaches may overlook the multi-modal data streams routinely generated during discovery and development, including permeability assays, metabolic stability studies, and formulation attributes. The objective of this predictive modeling article is to define a machine learning framework for estimating oral bioavailability from molecular, permeability, metabolism, and formulation features. The model is intended to support early ranking of compounds and formulation strategies rather than replace definitive pharmacokinetic studies. A gradient-boosted tree model would be trained on curated oral bioavailability measurements linked to chemical structures, in vitro permeability values, intrinsic clearance estimates, and formulation descriptors. Feature engineering would convert heterogeneous experimental and categorical information into a harmonized input vector suitable for interpretable prediction.

Conceptually, the model could predict oral bioavailability by learning non-linear relationships among molecular descriptors, epithelial transport surrogates, metabolic liability, and formulation class. It would also be expected to generate interpretable feature-attribution patterns and uncertainty estimates for risk-based decision making. A holistic, data-driven bioavailability model could accelerate candidate selection and formulation design in early drug development. Its greatest value would lie in integrating routinely available evidence into a single transparent prediction workflow.

To Cite This Article: Chen W, Zhang L. Machine Learning for Oral Bioavailability Prediction Using Molecular, Permeability, Metabolism, and Formulation Features. Pharmacophore. 2024;15(6):24-34. https://doi.org/10.51847/ECnRcM8N3p

Introduction

Oral bioavailability is a central determinant of whether a drug candidate can achieve sufficient systemic exposure after oral dosing, and poor pharmacokinetics can redirect or terminate otherwise promising programs. Early discovery teams have long used molecular rules and structure-derived descriptors to identify compounds likely to suffer from poor absorption or excessive clearance, but such rules are intentionally coarse and cannot fully represent the mechanisms governing oral exposure. Tools such as SwissADME formalized the rapid estimation of physicochemical and drug-likeness properties [1], while ADMET-oriented platforms extended structure-based prediction toward broader pharmacokinetic liabilities [2, 3]. However, oral bioavailability remains more difficult to predict than isolated properties because it emerges from linked processes of dissolution, permeation, metabolism, and systemic disposition.

A growing amount of biopharmaceutic data is generated before first-in-human studies, including Caco-2 or related permeability measures, microsomal or hepatocyte stability, and formulation design information. Permeability modeling studies have shown that epithelial transport surrogates can be learned from chemical and assay-derived features [4, 5], while recent work on PAMPA and Caco-2 prediction illustrates how assay-specific models can support broader absorption modeling [6, 7]. In parallel, machine learning approaches for intrinsic clearance and cytochrome P450 liability provide computable representations of first-pass metabolism [8-12]. Formulation-focused artificial intelligence frameworks further indicate that dosage form, excipient strategy, and delivery technology can be encoded for predictive development workflows [13].

Machine learning is attractive for oral bioavailability prediction because it can capture non-linear interactions between chemical structure, permeability, metabolism, and formulation context. Published models for human oral bioavailability have

Corresponding Author: Wei Chen; Department of AI-Based Pharmaceutical Engineering, Faculty of Pharmacy, School of Pharmaceutical Sciences, Peking University, Beijing, China. E-mail: wei.chen@outlook.com.

used cheminformatics descriptors and machine learning pipelines to estimate %F directly from curated datasets [14, 15], while more recent studies have explored graph neural networks and transfer learning for the same endpoint [16]. In animal and translational settings, models using chemical structure and in vitro or in silico inputs have also been applied to oral exposure and rat bioavailability prediction [17, 18]. These studies support a broader modeling premise: oral bioavailability can be treated as a composite ADME endpoint whose prediction should benefit from multi-modal feature integration.

This article proposes a predictive model that merges molecular descriptors, permeability features, metabolism features, and formulation descriptors into a unified machine learning workflow for oral bioavailability prediction. Gradient-boosted tree models are a suitable conceptual choice because they can handle mixed feature types, missingness patterns, and non-linear interactions without requiring a fully mechanistic description of every absorption and first-pass process [19, 20]. The proposed model would complement existing ADMET platforms rather than replace them, using structure-derived predictions, in vitro assay data, and formulation annotations as coordinated inputs [21–23]. Its primary output would be an interpretable estimate of expected oral bioavailability that could guide chemists, pharmacokineticists, and formulators during compound and formulation prioritization.

Background

Determinants of Oral Bioavailability

Oral bioavailability is determined by the fraction of dose released from the formulation, dissolved in gastrointestinal fluids, absorbed across the intestinal barrier, and escaping gut-wall and hepatic first-pass metabolism. Molecular descriptors such as lipophilicity, molecular weight, hydrogen-bonding capacity, polar surface area, and rotatable bonds influence dissolution and passive permeation, which explains why descriptor-based systems remain common in early screening [1]. Yet first-pass extraction depends heavily on metabolic susceptibility, including intrinsic clearance and CYP450 involvement, which can be modeled using clearance and enzyme-liability predictors [8, 10]. Formulation variables can further alter apparent absorption by changing solubilization, particle-size-dependent dissolution, and lipid-mediated absorption pathways, making oral bioavailability a multi-factorial endpoint rather than a direct molecular property [13, 24].

In Vitro and In Silico Predictors

Caco-2, MDCK, and PAMPA assays provide experimentally accessible permeability surrogates, but each captures only part of the intestinal absorption process and can be affected by assay protocol, transporter expression, and compound-specific ionization. Machine learning studies of Caco-2 and related permeability endpoints demonstrate that chemical descriptors and assay-derived data can be mapped to apparent permeability classes or values [4–6]. PAMPA-focused modeling similarly suggests that passive permeability may be represented computationally, but its relationship to in vivo absorption can be incomplete when active transport or metabolism dominates [7, 25]. For metabolism, microsomal and hepatocyte intrinsic clearance values offer useful first-pass indicators, while ML models for clearance and CYP interactions provide in silico complements when experimental data are incomplete [8–12].

Formulation as a Modifier of Bioavailability

Formulation can modify oral bioavailability by altering release rate, dissolution, supersaturation, intestinal solubilization, and lymphatic uptake. Food-effect modeling has shown that bioavailability can vary with administration conditions and formulation-dependent solubilization, reinforcing that a compound's structure alone cannot fully determine oral exposure [24]. Machine learning-directed formulation development provides a conceptual basis for encoding dosage form, excipient class, lipid-based formulation status, and solubility-enhancing technology as structured model inputs [13]. Web-based formulation design tools further show how categorical and material descriptors can be incorporated into artificial intelligence systems used by formulation scientists.

A biopharmaceutic framing also helps prevent the model from treating oral bioavailability as a purely structure-derived endpoint. The Biopharmaceutics Classification System links in vitro dissolution, gastrointestinal permeability, and in vivo bioavailability, making solubility and permeability mechanistically important variables for oral absorption modeling. Similarly, the Rule of 5 literature emphasizes that poor absorption or permeation becomes more likely when molecular size, lipophilicity, and hydrogen-bonding properties exceed drug-like ranges. Accordingly, the proposed feature set should encode dissolution- and permeability-relevant descriptors alongside formulation context, while allowing the learning algorithm to determine how these variables interact with clearance and first-pass extraction.

Existing Machine Learning Models for Bioavailability

Existing oral bioavailability models range from descriptor-based workflows to deep learning and transfer-learning approaches. A public KNIME workflow demonstrated how curated molecular descriptors could be used for human oral bioavailability prediction [14], and HobPre extended this paradigm with a dedicated small-molecule prediction framework [15]. Recent studies have explored graph neural networks, improved deep forest methods, and molecular-modification-oriented models for human oral bioavailability prediction [16, 26, 27]. Preclinical work has also shown that machine learning can be applied to rat bioavailability and mouse oral exposure, supporting the use of species-specific models when human data are limited [17, 18].

Integration with Physiologically-Based Pharmacokinetic Models

Physiologically based pharmacokinetic models offer mechanistic representations of absorption, distribution, metabolism, and excretion, but they require many input parameters that may be uncertain in early discovery. Machine learning can complement PBPK by rapidly estimating missing inputs, prioritizing compounds for more detailed simulation, or providing empirical screens before mechanistic modeling is justified [28, 29]. In silico prediction of PBPK input parameters has been used to support simplified exposure modeling after oral dosing, illustrating how computational estimates can feed mechanistic frameworks [30]. The proposed oral bioavailability model would therefore function as a practical bridge between empirical ADME prediction and later PBPK-based development decisions.

Model Development Overview

High-Level Prediction Pipeline

The proposed prediction pipeline would begin with a standardized compound record linking chemical structure, measured or predicted permeability, measured or predicted metabolism, and formulation information to an observed oral bioavailability value. Structure-derived descriptors would be calculated using cheminformatics workflows similar in purpose to those used in ADMETlab, SwissADME, and Chemprop-style molecular prediction systems [1, 20, 21]. Permeability and metabolism features would then be harmonized with formulation descriptors before being passed to a supervised machine learning model that outputs predicted human %F with an uncertainty interval. This pipeline would be designed to allow partial data availability, because early-stage compounds may have molecular descriptors before complete Caco-2, microsomal, hepatocyte, or formulation data are available.

Figure 1 presents the proposed formulation-aware machine learning architecture for integrating molecular, permeability, metabolism, and formulation evidence into an interpretable oral bioavailability prediction workflow.

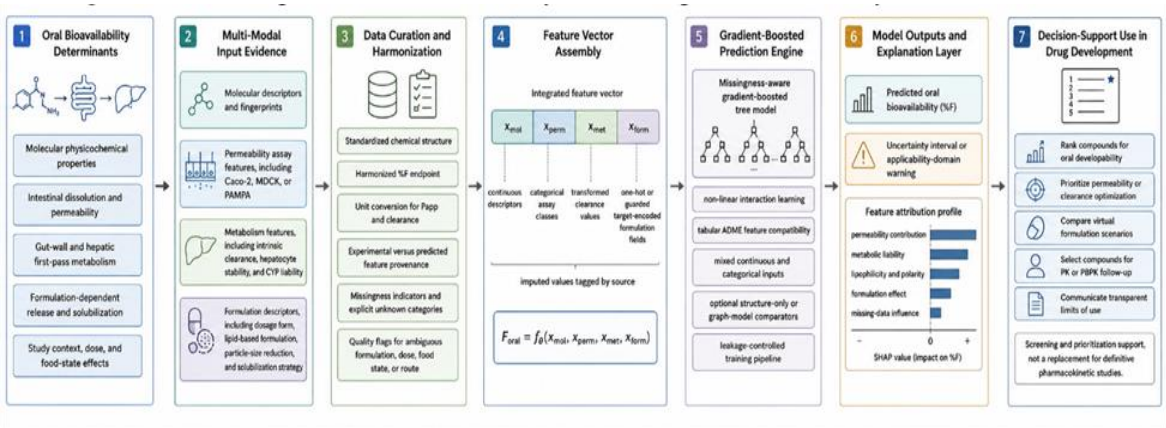

Figure 1. Integrated machine learning workflow for oral bioavailability prediction using molecular, permeability, metabolism, and formulation features

Core Input Feature Sets

The core molecular feature set would include 2D descriptors such as logP or logD, molecular weight, hydrogen bond donors and acceptors, topological polar surface area, ionization-related descriptors, aromaticity indicators, and rotatable bond counts. Permeability features would encode either numerical Papp values from Caco-2, MDCK, or PAMPA assays, or categorical high, moderate, and low permeability labels when numerical values are unavailable [4–7]. Metabolism features would include intrinsic clearance from microsomes or hepatocytes, CYP substrate or inhibitor flags, and metabolism-based extraction indicators informed by ML models of clearance and CYP450 liabilities [8–12]. Formulation features would include dosage form, solution or solid status, lipid-based formulation indicators, particle-size reduction, and solubilization strategy, reflecting the formulation variables emphasized in machine learning-directed formulation development [13, 24].

Design Principles

The model should be simple enough to deploy within discovery workflows while still broad enough to represent the major determinants of oral bioavailability. Public ADMET platforms show the value of accessible, structure-based prediction interfaces [2, 3, 21], while molecular machine learning packages and benchmarks show how modern descriptor, fingerprint, and graph-based representations can be operationalized for chemical property prediction [20, 31, 32]. The model should remain interpretable to project teams so that predicted bioavailability can be connected to actionable hypotheses, such as improving permeability, reducing clearance, or changing formulation strategy. It should also tolerate incomplete in vitro profiles by using imputation, missingness indicators, and optional in silico predictors rather than excluding compounds that lack every assay measurement.

Data Sources and Feature Engineering

Curation of Bioavailability Data

Human oral bioavailability values would be curated from peer-reviewed literature, public drug resources, and, where available, proprietary pharmacokinetic databases, with each record mapped to a standardized structure and formulation context. Existing bioavailability modeling studies demonstrate the importance of careful endpoint definition when building curated %F datasets [14, 15, 26, 27]. Public ADMET resources such as ADMETlab and admetSAR illustrate how diverse chemical and pharmacokinetic annotations can be organized into machine-readable prediction frameworks [2, 3, 21]. During curation, each %F value would be converted to a fractional or percentage endpoint consistently, while records with ambiguous routes, uncontrolled food effects, unclear formulation identity, or conflicting reports would be flagged rather than treated as equally reliable.

Molecular and Permeability Descriptors

Molecular descriptors would be computed from standardized 2D structures, using reproducible cheminformatics workflows that capture lipophilicity, size, hydrogen bonding, polarity, flexibility, and related properties. Such descriptors remain useful because benchmark studies and molecular property prediction frameworks show that structure-based representations can support a broad range of chemical and ADME endpoints [20, 31, 32]. Permeability features would be represented as continuous values when Caco-2, MDCK, or PAMPA Papp measurements are available, and as categorical classes when only qualitative assay interpretation is reported [4–7]. When measured permeability is missing, separate in silico permeability predictors could supply auxiliary features, but these values should be tagged as predicted rather than experimental to preserve provenance.

Metabolism and Formulation Encoding

Metabolism features would be harmonized by standardizing microsomal and hepatocyte intrinsic clearance units and deriving qualitative extraction-risk categories when mechanistically appropriate. Machine learning models for intrinsic clearance, metabolic drug interactions, and CYP450 inhibition or substrate liability provide a basis for filling or contextualizing metabolic variables when direct assays are unavailable [8–12]. Formulation descriptors would be encoded using one-hot or target-aware categorical representations for tablet, capsule, solution, suspension, lipid-based formulation, amorphous dispersion, particle-size reduction, and solubilization technology [13, 24]. An explicit unknown category would be retained for formulation and metabolism fields so that records are not discarded solely because legacy reports omit formulation details.

Table 1 defines the multi-modal feature architecture needed to convert heterogeneous molecular, ADME, formulation, and provenance data into a unified oral bioavailability prediction input.

Table 1. Multi-Modal Feature Architecture for Formulation-Aware Oral Bioavailability Prediction

Feature domain	Representative variables	Biopharmaceutical meaning	Encoding strategy	Decision relevance	Key risk if omitted
Molecular physicochemical profile	Molecular weight, logP/logD, hydrogen bond donors and acceptors, topological polar surface area, rotatable bonds, aromaticity, ionization-related descriptors	Captures chemical determinants of solubility, passive permeability, flexibility, polarity, and drug-likeness	Continuous descriptors, binned physicochemical categories, molecular fingerprints, optional graph-derived embeddings	Helps medicinal chemists identify whether low predicted %F is structurally driven	The model may reduce oral bioavailability to assay artifacts without recognizing intrinsic molecular limitations
Permeability evidence	Caco-2 Papp, MDCK Papp, PAMPA permeability, high/moderate/low permeability class, efflux indicators when available	Represents intestinal barrier crossing and passive or transporter-influenced absorption potential	Log-transformed continuous values, categorical permeability classes, missingness indicators, experimental versus predicted provenance flags	Guides whether compound optimization should prioritize epithelial transport, permeability improvement, or formulation-enabled absorption	Poorly permeable compounds may be incorrectly prioritized if structure-only descriptors appear favorable
Metabolic liability evidence	Microsomal intrinsic clearance, hepatocyte clearance, CYP substrate flags, CYP inhibition risk, predicted extraction-risk category	Represents gut-wall and hepatic first-pass loss after absorption	Standardized clearance units, transformed clearance values, binary CYP flags, ordinal extraction-risk classes	Helps distinguish absorption-limited from first-pass-metabolism-limited candidates	High-clearance compounds may be misclassified as orally developable when permeability is adequate
Formulation context	Solution, tablet, capsule, suspension, lipid-based system, amorphous dispersion, particle-size reduction, solubilization strategy, excipient class	Represents release, dissolution, supersaturation, solubilization, and delivery-dependent absorption modification	One-hot encoding, guarded target encoding, explicit unknown formulation category, formulation-family grouping	Enables virtual comparison of formulation strategies while holding molecular and ADME features constant	The model may incorrectly treat bioavailability as a fixed molecular property rather than a formulation-modifiable endpoint

Study and provenance metadata	Species, dose range, feeding state, route clarity, data source, assay protocol, endpoint quality flag, formulation reporting quality	Captures heterogeneity that may influence observed %F independent of compound biology	Source indicators, quality flags, stratification variables, exclusion or sensitivity-analysis labels	Supports reliability weighting, external validation design, and interpretation of uncertain records	Dataset noise may be learned as biological signal, reducing generalizability
Missingness and imputation indicators	Missing permeability, missing clearance, predicted rather than measured ADME input, unknown formulation, incomplete food-state metadata	Distinguishes absence of evidence from biological absence of a property	Binary missingness flags, imputation source labels, separate unknown categories, auxiliary prediction indicators	Allows early-stage compounds to remain usable without hiding uncertainty	The model may overtrust imputed or incomplete records and produce overconfident predictions

Predictive Model Architecture

Algorithm Choice and Rationale

A gradient-boosted tree model such as XGBoost would be a reasonable primary architecture because it can learn non-linear interactions among continuous descriptors, categorical formulation variables, and missingness-aware assay features. Gradient boosting is well suited to tabular ADME prediction, where input variables may include structure-derived descriptors, permeability values, clearance measurements, and formulation indicators rather than a single homogeneous feature type. Similar machine learning strategies have been applied to oral bioavailability, oral exposure, permeability, and broader molecular property prediction tasks, supporting their conceptual suitability for this endpoint [4–7, 14–19, 26, 27]. Deep learning or graph neural network models could be explored as secondary comparators, particularly when structure representation is expected to contribute information beyond engineered descriptors [16, 20, 32].

Feature Vector Assembly and Pre-processing

Feature vector assembly would begin by joining molecular descriptors, permeability records, metabolism features, and formulation annotations at the compound–study level. Missing permeability or metabolism values could be handled using model-based imputation, k-nearest-neighbor imputation, missingness indicators, or auxiliary predictions from permeability and ADMET models [2-7, 21]. Categorical formulation fields could be one-hot encoded or target-encoded with safeguards against leakage, while numerical variables such as intrinsic clearance or Papp could be transformed to reduce skew without claiming a specific empirical performance gain. The same preprocessing pipeline would be applied during training, validation, and deployment so that the predicted %F for a new compound reflects the same feature definitions used to develop the model. **Table 2** shows the structured feature vector assembly pipeline for %F prediction, including feature grouping, encoding strategies, and consistent handling of missing and heterogeneous pharmacokinetic data across all modeling stages.

Table 2. Feature vector assembly and preprocessing strategy for oral bioavailability (%F) modeling

Feature group	Example variables	Encoding / transformation	Missing data strategy	Notes
Molecular descriptors	MW, LogP, TPSA, HBD/HBA	Scaling (standardization or normalization)	Imputation or model-based prediction	Core physicochemical property set
Permeability features	Caco-2 Papp, PAMPA permeability	Log-transform or skew reduction	KNN imputation, ADMET model prediction	Key absorption driver
Metabolism features	Intrinsic clearance, CYP inhibition flags	Scaling; categorical encoding for enzyme interactions	Auxiliary ADMET model imputation	Reflects first-pass loss
Formulation annotations	Salt form, dosage form, release type	One-hot encoding or target encoding	Most-frequent imputation	Requires leakage control in target encoding
Study-level context	Species, assay type, protocol conditions	One-hot or embedding encoding	Missing category assignment	Harmonizes heterogeneous datasets
Missingness indicators	Missing Papp, missing clearance flags	Binary indicator variables	Not applicable	Helps model learn missing-data patterns

Output and Uncertainty Quantification

The primary output would be a predicted oral bioavailability value expressed as %F, accompanied by an uncertainty interval or prediction band for decision support. Uncertainty could be estimated through quantile regression, conformal prediction, ensemble dispersion, or applicability-domain analysis, especially because oral bioavailability records can vary by species, formulation, dose, and study conditions. Published work on ADMET prediction, PBPK input estimation, and pharmacokinetic modeling highlights the need to represent uncertainty when computational predictions inform development choices [28–30, 33]. The model output should therefore be interpreted as a ranked, risk-informed estimate that guides compound or formulation prioritization, not as a definitive substitute for in vivo pharmacokinetic evaluation.

Handling Data Heterogeneity and Imbalanced Observations

Multi-Source Data Integration and Adjustment

Oral bioavailability datasets would be heterogeneous because values may originate from different laboratories, species, analytical methods, dose levels, feeding states, and formulation descriptions. Source-specific dummy variables or hierarchical adjustment terms could help the model distinguish biological signal from study-origin effects, particularly when public and proprietary data are combined [14, 15]. Similar concerns arise in permeability and clearance modeling, where assay protocol and experimental context can influence apparent values even when the underlying chemical structure is unchanged [6, 8]. A practical formulation-aware model should therefore retain provenance fields and quality flags rather than collapsing all observations into a single unqualified endpoint.

Handling a Small Number of Compounds with Full Assays

Only a subset of compounds would be expected to have complete molecular, permeability, metabolism, and formulation profiles, especially in early discovery. Semi-supervised learning, multi-task learning, or staged modeling could allow the framework to learn from compounds that have only molecular descriptors while still using richer assay profiles when available [20, 22]. Transfer learning and graph-based representations may also help extract structural information from larger chemical datasets before fine-tuning on oral bioavailability labels [16, 32]. In this design, incomplete records would remain useful because missingness itself may reflect the stage of development, assay priority, or historical data availability rather than irrelevance.

Data Augmentation with In Silico Predictors

Data augmentation could use separate in silico models to estimate missing permeability, clearance, CYP liability, or general ADMET properties when experimental values are unavailable. Caco-2 and PAMPA prediction models provide plausible auxiliary inputs for intestinal permeation risk [5–7, 25], while CYP and intrinsic-clearance predictors can supply metabolism-oriented features for compounds lacking direct microsomal or hepatocyte data [8–12]. ADMET platforms such as ADMETlab, admetSAR, SwissADME, and Deep-PK further illustrate how computationally predicted properties can be assembled into a broader pharmacokinetic feature set [1–3, 21, 22]. These augmented values should be clearly labeled as predicted features so that the final model can learn their uncertainty and avoid treating them as equivalent to measured assay data.

Table 3 shows the in silico data augmentation strategies used to impute missing permeability, metabolic stability, CYP liability, and broader ADMET properties through predictive computational models and integrated pharmacokinetic platforms.

Table 3. In silico data augmentation strategies for missing ADMET and pharmacokinetic properties

Property type	Missing experimental data	In silico models / tools	Output features used in ML pipeline	Key notes
Intestinal permeability	Caco-2, PAMPA assay values	Caco-2 prediction models, PAMPA QSAR models	Predicted permeability (e.g., Papp, logPapp)	Proxy for intestinal absorption; may vary across model systems
Passive diffusion	Human or in vivo absorption data	Physicochemical + ML permeability models	Absorption likelihood scores	Captures membrane diffusion tendency but not active transport
Metabolic stability	Microsomal/hepatocyte stability data	Intrinsic clearance prediction models	Predicted Clint, half-life proxies	Reflects hepatic metabolism rate; high uncertainty in extrapolation
CYP liability	CYP450 inhibition/induction assay results	CYP interaction prediction models (QSAR/deep learning)	Probability of CYP inhibition/induction per isoform	Critical for drug–drug interaction risk estimation
Hepatic clearance	In vivo clearance measurements	PK/clearance prediction models	Predicted systemic clearance (CL)	Often derived from compound structure and in vitro surrogates
General ADMET profile	Multi-assay experimental ADMET panels	ADMET platforms (ADMETlab, admetSAR, SwissADME, Deep-PK)	Composite ADMET descriptors	Aggregates absorption, distribution, metabolism, toxicity features

Model Interpretability and Biopharmaceutical Insight

Global and Local SHAP Analysis

Model interpretation would be centered on global and local attribution methods that rank the contribution of permeability, clearance, lipophilicity, polarity, ionization, and formulation descriptors to predicted oral bioavailability. SHAP-style explanations are especially useful for tabular ADME models because they can identify whether a prediction is mainly driven by poor permeability, high metabolic liability, or formulation limitations rather than by a single molecular rule [19, 20]. For an individual compound, a waterfall explanation could show how a favorable permeability class offsets moderate clearance,

or how a lipid-based formulation indicator increases the predicted likelihood of useful exposure [13, 24]. Such explanations would help medicinal chemists and formulation scientists convert a black-box prediction into a testable optimization hypothesis.

Translating Model Logic into Design Guidance

The interpretability layer should translate model behavior into design guidance by identifying which intervention would be expected to improve oral bioavailability for a given scaffold or formulation context. For example, if predicted first-pass extraction dominates the attribution pattern, chemistry efforts aimed only at increasing lipophilicity may be less rational than reducing intrinsic clearance or avoiding CYP-mediated liability [8, 10]. Conversely, if permeability and dissolution-related descriptors dominate, the model could prioritize permeability improvement, particle-size control, or solubilization strategies before recommending metabolism-focused changes [4, 7, 13]. This logic would make the model a decision-support system rather than only a numerical predictor.

Figure 2 summarizes how prediction outputs, attribution patterns, expert review, and new experimental evidence would be linked in an iterative decision-support loop.

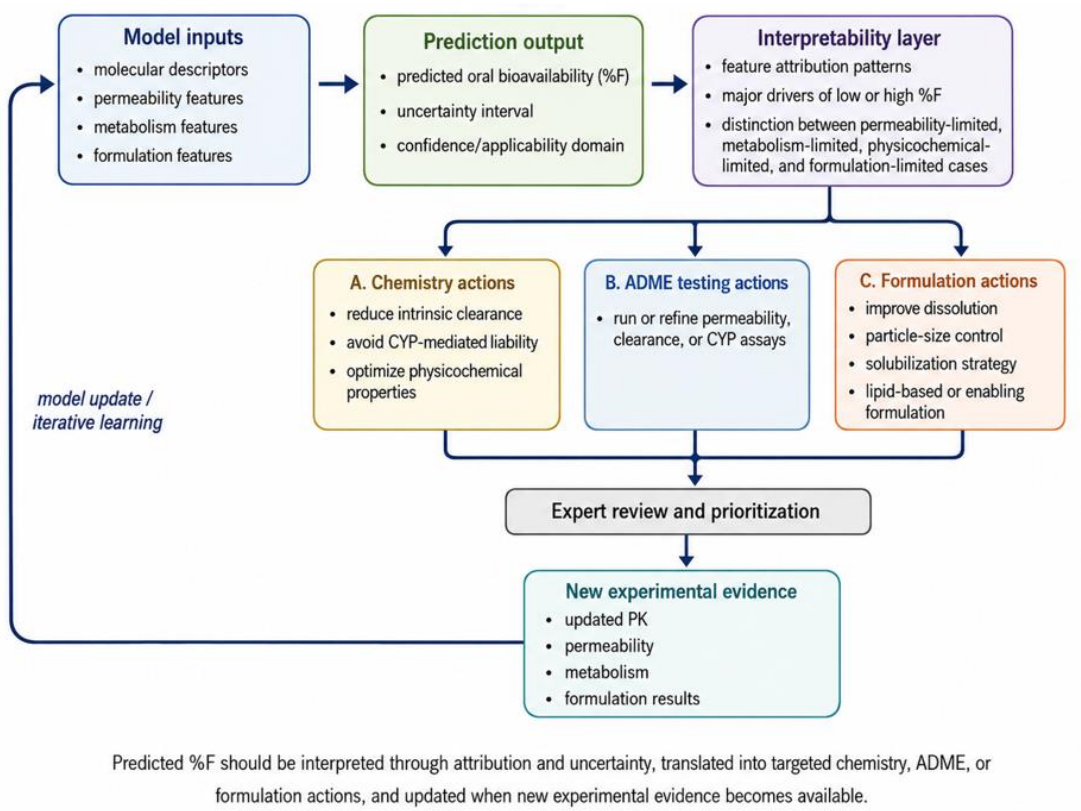

Figure 2. Interpretation-to-action loop for using oral bioavailability predictions in compound and formulation decision making. Caption: Predicted %F should be interpreted through attribution and uncertainty, translated into targeted chemistry, ADME, or formulation actions, and updated when new experimental evidence becomes available.

Integration Into Drug Development Workflow

Early Discovery Triage

In early discovery, the model could be deployed as a web-based tool or compound-registration plugin that returns a predicted oral bioavailability estimate when a new structure is proposed. Existing ADMET and formulation platforms demonstrate that browser-accessible prediction environments can support rapid project decisions when they combine standardized inputs, transparent outputs, and practical interpretability [2, 3, 21]. The model would be most useful when it presents not only a predicted %F but also a reasoned attribution summary and an applicability-domain warning. This workflow would allow project teams to rank compound series before committing to more expensive permeability, metabolic stability, formulation, or animal pharmacokinetic studies.

Formulation Candidate Selection

For formulation scientists, the same model could be used to compare virtual formulation scenarios by changing dosage form, lipid-based formulation status, solubilization technology, or particle-size category while keeping molecular, permeability, and metabolism features fixed. Machine learning-directed formulation development supports this kind of scenario analysis because

formulation attributes can be encoded as model inputs and evaluated before extensive prototype manufacture [13]. The general prediction function can be written as $F_{oral} = f_{\theta(x_{mol},x_{perm},x_{met},x_{form})}$, where F_{oral} is the predicted oral bioavailability, f_θ is the trained machine learning model, x_{mol} represents molecular descriptors, x_{perm} represents permeability features, x_{met} represents metabolism features, and x_{form} represents formulation features. By changing only x_{form}, the model would allow formulators to estimate how an enabling formulation might alter predicted oral bioavailability without implying that the predicted change is an experimentally confirmed uplift.

Evaluation Strategy

Prediction Accuracy and Generalization

Evaluation should examine whether the model generalizes across compounds, scaffolds, sources, species contexts, and formulation classes without reporting unsupported performance values. Cross-validation, temporal validation, scaffold-based splitting, and external validation would each test a different failure mode, and similar validation principles are used across molecular property prediction and pharmacokinetic modeling studies [20, 31–33]. Regression metrics such as RMSE, MAE, and R^2 could be prespecified, while classification-style summaries could evaluate whether the model distinguishes lower- from higher-bioavailability compounds. The key requirement is that all reported metrics would come only from actual validation, not from assumed or illustrative results.

Benchmarking Against Existing Tools

Benchmarking should compare the proposed multi-modal model against simpler molecular rules, structure-only ADMET tools, and published oral bioavailability models on the same validation sets. Descriptor-based oral bioavailability workflows, HobPre, graph neural network models, deep forest approaches, and molecular-modification-oriented predictors provide relevant conceptual comparators [14–16, 26, 27]. Broader ADMET systems such as SwissADME, ADMETlab, admetSAR, and Deep-PK could also serve as practical baselines for structure-derived pharmacokinetic risk estimation [1–3, 21, 22]. A fair benchmark would determine whether adding permeability, metabolism, and formulation features improves decision support beyond what can be inferred from molecular structure alone.

Prospective Virtual Screening

Prospective virtual screening would test whether the model can support real project decisions before new oral pharmacokinetic data are available. A time-split design could train the model on earlier compounds and evaluate predictions on later compounds whose measured bioavailability became available after model development, mirroring real discovery uncertainty [15, 16]. The same concept could be applied to formulation strategy by predicting relative benefit for alternative dosage forms or solubilization approaches before prototype testing [13, 24]. This evaluation would be especially important because retrospective performance can overstate practical utility when closely related compounds or shared assay sources appear in both training and test data.

Table 4 provides a deployment-readiness framework linking model validation, uncertainty estimation, interpretability, benchmarking, and governance to practical oral drug-development decisions.

Table 4. Validation, Interpretability, and Deployment Readiness Framework for Oral Bioavailability Prediction

Evaluation dimension	Main question addressed	Recommended analytical approach	Failure mode detected	Interpretation for drug-development decision making
Random cross-validation	Does the model learn a reproducible signal within the curated dataset?	K-fold cross-validation with prespecified RMSE, MAE, R^2, and calibration summaries	Overfitting to noisy endpoint records or unstable preprocessing	Useful for initial model checking but insufficient as evidence of deployment readiness
Scaffold-based validation	Does the model generalize beyond closely related chemical series?	Chemical scaffold split or cluster-based split separating related chemotypes	Memorization of analog-series patterns rather than transferable ADME logic	Determines whether predictions can support novel compound-series prioritization
Temporal validation	Would the model have predicted later compounds from earlier data?	Train on earlier records and test on later records based on discovery or publication date	Retrospective optimism caused by leakage from future-like records	Approximates real project use when teams must rank compounds before new PK data exist
External dataset validation	Does the model transfer across sources, laboratories, and reporting standards?	Independent validation using external public, institutional, or proprietary data	Source-specific bias, assay-protocol dependency, and poor cross-dataset robustness	Required before the model can be trusted outside the development dataset
Formulation-stratified validation	Does performance hold across solution, solid, lipid-based, and solubilization-enabled formulations?	Subgroup performance analysis by formulation class and reporting quality	False confidence when formulation effects are underreported or unevenly represented	Determines whether the model can support formulation scenario comparison

Uncertainty and applicability-domain assessment	Does the model know when it is extrapolating?	Conformal prediction, quantile regression, ensemble dispersion, nearest-neighbor chemical-domain analysis	Overconfident prediction for unusual scaffolds, sparse formulation classes, or incomplete ADME profiles	Converts predictions into risk-informed estimates rather than unsupported point values
Interpretability review	Are predictions explainable in biopharmaceutical terms?	Global and local SHAP analysis, feature interaction inspection, attribution review by domain experts	Spurious reliance on provenance, missingness, or non-causal categorical features	Helps chemists, pharmacokineticists, and formulators identify actionable optimization hypotheses
Benchmark comparison	Does the multi-modal model add value beyond structure-only or rule-based approaches?	Compare against molecular rules, structure-only ADMET tools, published oral bioavailability models, and simpler baseline regressors	Added complexity without improved decision support	Justifies whether permeability, metabolism, and formulation features are worth collecting
Prospective decision-impact testing	Does the model improve compound or formulation prioritization before new PK data are generated?	Use predictions to rank candidates or formulation scenarios, then compare against later experimental outcomes	Retrospective accuracy that fails to improve real development decisions	Provides the strongest evidence that the model is practically useful rather than only statistically accurate
Governance and use-boundary review	Are predictions deployed with appropriate limits, documentation, and human oversight?	Model card, data provenance report, version control, audit trail, expert review checkpoint, retraining trigger	Uncontrolled use as a definitive PK substitute or use outside validated domain	Positions the model as transparent screening support, not a replacement for in vivo PK or PBPK evaluation

Limitations

Data Availability and Quality

Human oral bioavailability data are scarce, noisy, and often reported without complete formulation, food-state, dose, or study-design metadata. Existing bioavailability and oral-exposure models demonstrate that curated datasets can support useful prediction, but they also depend strongly on endpoint harmonization and data-quality control [14, 15, 17, 18]. In vitro permeability and intrinsic clearance values may also vary across laboratories and assay protocols, which can introduce uncertainty before the model even sees the data [6, 8]. Therefore, the model's most reliable use would be comparative prioritization within a defined chemical or formulation context rather than universal prediction across every oral drug class.

1Biological Complexity beyond Features

The proposed feature set would not fully capture active transport, intestinal metabolism, gut microbiome transformation, bile effects, food effects, enterohepatic cycling, or dose-dependent solubility and permeability. Food-effect modeling and PBPK-oriented studies show that oral exposure can depend on physiological context and mechanistic parameters that are not always available in early discovery datasets [24, 28–30]. Even strong molecular, permeability, metabolism, and formulation features may fail for compounds dominated by transporter saturation, unusual regional absorption, or complex precipitation and redissolution behavior. The model should therefore be positioned as a screening and prioritization tool that complements PBPK simulation and definitive pharmacokinetic studies rather than replacing them.

Conclusion

A formulation-aware machine learning model for oral bioavailability prediction would integrate the major evidence streams that shape oral exposure. By combining molecular descriptors, permeability measurements or predictions, metabolism indicators, and formulation descriptors, the model could provide a more holistic estimate of expected %F than structure-only rules.

The model's major strength would be its ability to convert heterogeneous discovery data into a single interpretable decision-support output. Attribution methods could help teams understand whether low predicted oral bioavailability is more likely driven by permeability, metabolic extraction, physicochemical limitations, or formulation opportunity.

Important challenges would remain, including sparse human bioavailability labels, inconsistent formulation reporting, noisy in vitro assays, and biological mechanisms that are difficult to encode in simple tabular features. External and prospective validation would be essential before the model could be used confidently for candidate nomination or formulation prioritization.

The field would benefit from open benchmark datasets that include not only structures and %F values but also permeability, metabolism, and formulation annotations. Integration into pharmaceutical workflows would encourage broader adoption, provided that predictions are delivered with uncertainty, interpretability, and clear limits of use.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. 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.
2. Dong J, Wang NN, Yao ZJ, Zhang L, Cheng Y, Ouyang D, et al. ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. J Cheminform. 2018;10(1):29.
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. Wang NN, Huang C, Dong J, Yao ZJ, Zhu MF, Deng ZK, et al. Predicting human intestinal absorption with modified random forest approach: a comprehensive evaluation of molecular representation, unbalanced data, and applicability domain issues. RSC Adv. 2017;7(31):19007-18.
5. Acuña-Guzman V, Montoya-Alfaro ME, Negrón-Ballarte LP, Solis-Calero C. A machine learning approach for predicting Caco-2 cell permeability in natural products from the biodiversity in Peru. Pharmaceuticals. 2024;17(6):750.
6. Steinbauer FF, Lehr T, Reichel A. Exploring the potential of adaptive, local machine learning in comparison to the prediction performance of global models: a case study from Bayer's Caco-2 permeability database. J Chem Inf Model. 2024;64(24):9163-72.
7. Rácz A, Vincze A, Volk B, Balogh GT. Extending the limitations in the prediction of PAMPA permeability with machine learning algorithms. Eur J Pharm Sci. 2023;188:106514.
8. Rodriguez-Perez R, Trunzer M, Schneider N, Faller B, Gerebtzoff G. Multispecies machine learning predictions of in vitro intrinsic clearance with uncertainty quantification analyses. Mol Pharm. 2022;20(1):383-94.
9. Keefer CE, Chang G, Di L, Woody NA, Tess DA, Osgood SM, et al. The comparison of machine learning and mechanistic in vitro–in vivo extrapolation models for the prediction of human intrinsic clearance. Mol Pharm. 2023;20(11):5616-30.
10. Wang NN, Wang XG, Xiong GL, Yang ZY, Lu AP, Chen X, et al. Machine learning to predict metabolic drug interactions related to cytochrome P450 isozymes. J Cheminform. 2022;14(1):23.
11. Plonka W, Stork C, Šícho M, Kirchmair J. CYPlebrity: machine learning models for the prediction of inhibitors of cytochrome P450 enzymes. Bioorg Med Chem. 2021;46:116388.
12. Nguyen-Vo TH, Trinh QH, Nguyen L, Nguyen-Hoang PU, Nguyen TN, Nguyen DT, et al. iCYP-MFE: identifying human cytochrome P450 inhibitors using multitask learning and molecular fingerprint-embedded encoding. J Chem Inf Model. 2021;62(21):5059-68.
13. Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C. Machine learning directed drug formulation development. Adv Drug Deliv Rev. 2021;175:113806.
14. Falcón-Cano G, Molina C, Cabrera-Pérez MÁ. ADME prediction with KNIME: development and validation of a publicly available workflow for the prediction of human oral bioavailability. J Chem Inf Model. 2020;60(6):2660-7.
15. Wei M, Zhang X, Pan X, Wang B, Ji C, Qi Y, et al. HobPre: accurate prediction of human oral bioavailability for small molecules. J Cheminform. 2022;14(1):1.
16. Ng SS, Lu Y. Evaluating the use of graph neural networks and transfer learning for oral bioavailability prediction. J Chem Inf Model. 2023;63(16):5035-44.
17. Schneckener S, Grimbs S, Hey J, Menz S, Osmers M, Schaper S, et al. Prediction of oral bioavailability in rats: transferring insights from in vitro correlations to (deep) machine learning models using in silico model outputs and chemical structure parameters. J Chem Inf Model. 2019;59(11):4893-905.
18. Mughal H, Wang H, Zimmerman M, Paradis MD, Freundlich JS. Random forest model prediction of compound oral exposure in the mouse. ACS Pharmacol Transl Sci. 2021;4(1):338-43.
19. Peteani G, Huynh MT, Gerebtzoff G, Rodríguez-Pérez R. Application of machine learning models for property prediction to targeted protein degraders. Nat Commun. 2024;15(1):5764.
20. Heid E, Greenman KP, Chung Y, Li SC, Graff DE, Vermeire FH, et al. Chemprop: a machine learning package for chemical property prediction. J Chem Inf Model. 2023;64(1):9-17.
21. 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.
22. 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.
23. de Sá AG, Long Y, Portelli S, Pires DE, Ascher DB. toxCSM: comprehensive prediction of small molecule toxicity profiles. Brief Bioinform. 2022;23(5):bbac337.
24. Bennett-Lenane H, Griffin BT, O'Shea JP. Machine learning methods for prediction of food effects on bioavailability: a comparison of support vector machines and artificial neural networks. Eur J Pharm Sci. 2022;168:106018.
25. Brocke SA, Degen A, MacKerell AD Jr, Dutagaci B, Feig M. Prediction of membrane permeation of drug molecules by combining an implicit membrane model with machine learning. J Chem Inf Model. 2018;59(3):1147-62.

26. Yang Q, Fan L, Hao E, Hou X, Deng J, Xia Z, et al. Construction of an oral bioavailability prediction model based on machine learning for evaluating molecular modifications. J Pharm Sci. 2024;113(5):1155-67.
27. Ma L, Yan Y, Dai S, Shao D, Yi S, Wang J, et al. Research on prediction of human oral bioavailability of drugs based on improved deep forest. J Mol Graph Model. 2024;133:108851.
28. Naga D, Parrott N, Ecker GF, Olivares-Morales A. Evaluation of the success of high-throughput physiologically based pharmacokinetic (HT-PBPK) modeling predictions to inform early drug discovery. Mol Pharm. 2022;19(7):2203-16.
29. Chou WC, Lin Z. Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. Toxicol Sci. 2023;191(1):1-4.
30. Kamiya Y, Handa K, Miura T, Yanagi M, Shigeta K, Hina S, et al. In silico prediction of input parameters for simplified physiologically based pharmacokinetic models for estimating plasma, liver, and kidney exposures in rats after oral doses of 246 disparate chemicals. Chem Res Toxicol. 2021;34(2):507-13.
31. 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.
32. Wieder O, Kohlbacher S, Kuenemann M, Garon A, Ducrot P, Seidel T, et al. A compact review of molecular property prediction with graph neural networks. Drug Discov Today Technol. 2020;37:1-2.
33. Obrezanova O, Martinsson A, Whitehead T, Mahmoud S, Bender A, Miljkovic F, et al. Prediction of in vivo pharmacokinetic parameters and time–exposure curves in rats using machine learning from the chemical structure. Mol Pharm. 2022;19(5):1488-504.