Pharmacophore

ISSN-2229-5402

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

MECHANISTIC ACCOUNTABILITY IN HYBRID PHYSIOLOGICALLY BASED PHARMACOKINETIC AND MACHINE-LEARNING MODELS FOR DRUG DEVELOPMENT

Robert Miller[1]*, Laura Schmidt[2], James Walker[3], Elizabeth Taylor[4]

1. ***Department of Hybrid PBPK and Machine-Learning Modeling, College of Pharmacy, Pennsylvania State University, University Park, United States.***
2. ***Department of Mechanistic Accountability in PK Models, Faculty of Pharmacy, Virginia Tech, Blacksburg, United States.***
3. ***Department of Drug Development and Translational PK, Faculty of Pharmacy, University of Edinburgh, Edinburgh, United Kingdom.***
4. ***Department of PBPK-ML Integration and Validation, Faculty of Pharmaceutical Sciences, University of Guelph, Guelph, Canada.***

ARTICLE INFO

Received:
19 February 2025
Received in revised form:
21 April 2025
Accepted:
24 April 2025
Available online:
28 April 2025

Keywords: Physiologically based pharmacokinetics, Machine learning, Hybrid modeling, Mechanistic accountability, Model credibility, Uncertainty propagation

ABSTRACT

Hybrid physiologically based pharmacokinetic and machine-learning models offer a potentially valuable means of combining physiological organization with data-adaptive estimation. Their scientific interpretation is difficult, however, because predictive performance does not reveal whether a learned component represents a biological quantity, compensates for structural error, exploits dataset-specific associations, or remains reliable under extrapolation. This article develops mechanistic accountability as an original modeling principle for hybrid PBPK–machine-learning systems in drug development. Mechanistic accountability is defined as a traceable, context-specific justification linking each learned quantity to its data provenance, semantic meaning, interface with PBPK structure, effect on mechanistic states, identifiability, propagated uncertainty, biological plausibility, validation evidence, comparator performance, and permissible decision influence. The proposed construct distinguishes PBPK structure from physiological truth, learned inputs from measured parameters, prediction from mechanistic explanation, explainability from accountability, and model confidence from calibrated uncertainty. It further separates four possible machine-learning roles: estimating PBPK inputs, learning parameter or covariate functions, correcting mechanistic discrepancy, and directly predicting pharmacokinetic outputs. Each role creates different evidentiary obligations and failure modes. The article argues that hybrid-model evaluation should integrate software and interface verification, identifiability analysis, uncertainty propagation, biological challenge testing, external predictive assessment, and comparison with simpler alternatives. Qualification should remain bounded to a defined drug-development question, population, scenario, model influence, and consequence of error. The proposed framework is conceptual rather than empirically validated and does not establish regulatory acceptance, clinical utility, or deployment readiness. Its principal implication is that hybrid models should be judged not by a single performance measure but by whether their learned and mechanistic elements can sustain a transparent, proportionate, and scientifically defensible claim.

To Cite This Article: Miller R, Schmidt L, Walker J, Taylor E. Mechanistic Accountability in Hybrid Physiologically Based Pharmacokinetic and Machine-Learning Models for Drug Development. Pharmacophore. 2025;16(2):98-107. https://doi.org/10.51847/1Ae49RLXIg

Introduction

Mechanistic models, regulatory uses of physiologically based pharmacokinetic analysis, and expanding applications of machine learning collectively place hybrid modeling within a consequential pharmaceutical decision environment rather than a methodologically neutral computational setting [1–3]. PBPK models organize drug disposition through representations of physiological spaces, blood flows, tissue partitioning, binding, metabolism, transport, and elimination. Machine-learning models, by contrast, can estimate complex associations from chemical, biological, preclinical, and clinical data without requiring every relationship to be encoded as an explicit physiological equation. Combining these approaches may expand the

Corresponding Author: Robert Miller; Department of Hybrid PBPK and Machine-Learning Modeling, College of Pharmacy, Pennsylvania State University, University Park, United States. E-mail: robert.miller@psu.edu.

range of inputs that can be estimated, improve the representation of nonlinear covariate effects, or provide predictive functions where conventional parameterization is difficult. Yet the combination also creates a basic interpretive problem: once a learned component changes a PBPK parameter, state trajectory, residual pattern, or exposure prediction, it may become unclear which part of the resulting claim is supported by physiology, which part is supported by data association, and which part remains conditional on modeling assumptions.

This ambiguity matters because the evidentiary burden assigned to a model is not independent of the drug-development question or of the influence given to its output [4]. A hybrid model used to generate hypotheses about uncertain absorption mechanisms does not require the same evidence as a model intended to replace an interaction study, support extrapolation to an unobserved population, or materially influence a clinical-development decision. Nevertheless, hybrid systems are often described through aggregate predictive performance, selected goodness-of-fit displays, or a broad assertion that mechanistic structure improves interpretability. Such evidence may be informative, but it does not establish whether the learned component has an appropriate biological meaning, whether its uncertainty has been transmitted through the PBPK equations, whether the mechanistic parameters remain identifiable, or whether the same prediction could have been obtained from a simpler and less assumption-dependent model.

The unresolved problem is therefore not simply how to combine PBPK and machine learning technically. It is how to preserve a defensible chain of interpretation after the combination has occurred. A PBPK scaffold cannot automatically confer mechanistic status on a learned function merely because that function appears inside a physiological model. Conversely, the use of a learned component does not necessarily make the entire model opaque or scientifically uninterpretable. The relevant question is where learning occurs, what quantity is learned, how that quantity enters the mechanistic equations, which states and outputs it changes, and what evidence is available to justify those changes. Without this information, the terms *hybrid*, *mechanistic*, and *interpretable* risk becoming broad architectural labels rather than testable scientific properties.

This article proposes mechanistic accountability as an organizing principle for hybrid PBPK–machine-learning models in drug development. Mechanistic accountability is defined here as a traceable, context-specific justification linking each learned quantity to its provenance, semantic definition, PBPK interface, effect on mechanistic states, identifiability, propagated uncertainty, biological plausibility, validation evidence, simpler-model comparator, and permissible pharmaceutical claim. The construct is proposed as a methodological and conceptual contribution requiring future empirical evaluation. It is not presented as a formal regulatory standard, a universally applicable architecture, or evidence that hybrid models are intrinsically superior. Its purpose is to establish how claims about prediction, mechanism, extrapolation, and decision support should be bounded when physiological structure and machine learning are combined.

Why Hybrid Models Need Mechanistic Accountability

Post hoc explanations and predictive performance do not independently establish that a high-consequence model is intrinsically interpretable, scientifically valid, or suitable for the decision in which it may be used [5–7]. This distinction is especially important for hybrid PBPK–machine-learning systems because their apparent interpretability can arise from the visibility of the PBPK component even when the behavior of the learned component remains weakly understood. A feature-importance plot may indicate which molecular descriptors influenced a learned clearance estimate, but it does not establish that those descriptors represent the biological determinants of clearance. Similarly, agreement between predicted and observed concentrations does not show that the correct physiological mechanism produced the agreement. Mechanistic accountability therefore extends beyond explanation. It asks whether the learned component has a defined scientific role, whether that role is supported by appropriate evidence, and whether downstream claims remain proportionate to what the model can actually establish.

Responsible use of machine learning requires attention to data provenance, evaluation design, intended environment, monitoring, and the allocation of responsibility among relevant stakeholders [8]. Within a hybrid PBPK system, these responsibilities become component-specific. The developer of a learned input model must document the origin, suitability, and limitations of the training labels. The PBPK modeler must show how the learned quantity is transformed, scaled, bounded, and inserted into the physiological equations. The pharmacometrician or clinical pharmacologist must determine whether the resulting state trajectories and exposure predictions remain plausible for the population, regimen, and scenario being considered. Decision-makers must then interpret the prediction according to its residual uncertainty and consequence of error. Mechanistic accountability is therefore not a claim that every learned relationship must be causal. It is a requirement that associative, predictive, mechanistic, and decision-support claims remain visibly distinct.

Machine-learning applications in pharmacometrics also introduce challenges involving extrapolation, data adequacy, interpretability, validation, and integration with established modeling practice [9]. These challenges are amplified when errors can pass sequentially from data to a learned quantity, from that quantity to a PBPK parameter or function, and from the affected mechanistic states to a reported prediction. A model may therefore exhibit acceptable aggregate performance while failing for a particular chemical class, population, dose range, disease state, or interaction mechanism. The proposed accountability principle addresses this problem through six linked evidence domains: the scientific purpose of hybridization; the meaning and provenance of the learned quantity; the integrity of its interface with PBPK structure; the identifiability and uncertainty of the resulting model; the biological and predictive evaluation of its outputs; and the bounded context in which those outputs may influence development. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why hybrid models need mechanistic accountability within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for why hybrid models need mechanistic accountability

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Purpose of hybridization	Why is a learned component being added to the PBPK model?	A component should address a defined limitation, missing input, nonlinear relationship, discrepancy, or prediction task.	Connects architectural complexity to a specific pharmaceutical-science need.	Hybridization may be treated as inherently innovative or superior without demonstrating incremental value.	Added complexity is justified only relative to a defined question and suitable alternatives.
Data provenance and suitability	What data generated the learned relationship, and are those data appropriate for the intended application?	Training labels, assay conditions, populations, chemical space, and missingness patterns determine what the component can learn.	Establishes the evidentiary basis and domain of applicability of learned quantities.	Available data may be mistaken for representative, biologically meaningful, or decision-suitable data.	Data availability does not establish data suitability or transportability.
Semantic definition of the learned quantity	What biological, physicochemical, statistical, or residual quantity is being estimated?	A learned parameter requires a clear definition, unit, timescale, and relationship to observations.	Prevents an algorithmic output from being treated as a measured or causal biological property.	Latent or proxy outputs may be given stronger mechanistic interpretations than their labels support.	A learned quantity is mechanistically interpretable only to the extent that its meaning is defined and supported.
PBPK–machine-learning interface	How does the learned quantity enter the mechanistic model?	Interface integrity requires explicit transformations, units, timing, directionality, constraints, and software implementation.	Makes the transfer from association to mechanistic state evolution auditable.	Silent transformations or semantic mismatches can generate plausible but scientifically invalid predictions.	A transparent interface is necessary but does not validate the learned quantity itself.
Identifiability	Can the added parameters, functions, or latent quantities be distinguished from alternative explanations?	Structural and practical identifiability govern whether fitted components can sustain mechanistic interpretation.	Restricts claims about parameter meaning and causal or physiological explanation.	Flexible learned components may compensate for several mechanistic errors simultaneously.	Good output fit does not resolve non-identifiability.
Uncertainty propagation	How does uncertainty in data, learned inputs, parameters, and structure affect states and predictions?	Upstream errors can be amplified, attenuated, or correlated through PBPK equations.	Separates calibrated predictive uncertainty from algorithmic confidence.	Point predictions may conceal substantial uncertainty or unsupported extrapolation.	Uncertainty claims are conditional on the sources represented and the dependencies modeled.
Biological plausibility	Do parameter values, state trajectories, and perturbation responses remain compatible with established physiology and pharmacology?	Physiological ranges, conservation relations, directional responses, and perturbation evidence provide challenge conditions.	Tests whether acceptable predictions arise for scientifically credible reasons.	A model may reproduce observations through biologically incorrect compensation.	Biological plausibility is necessary for strong mechanistic claims but is not sufficient for validity.
Predictive validation	Does the model predict independent observations relevant to its intended use?	External or withheld evidence should reflect the compounds, populations, regimens, and scenarios of interest.	Assesses prospective adequacy within a defined application domain.	Retrospective fit or random data splitting may overstate generalizability.	Predictive validity is local to the evaluated domain and does not establish mechanism or clinical utility.
Simpler-model comparison	Does the hybrid architecture provide defensible value beyond a simpler PBPK, empirical, or machine-learning model?	Matched comparisons reveal whether added flexibility improves calibration, robustness, extrapolation, or decision relevance.	Makes complexity an empirical and context-dependent choice rather than a default.	A hybrid model may appear advantageous because comparators were absent or evaluated differently.	Hybrid complexity is not justified by novelty or one favorable metric alone.
Context-of-use boundary	What decision is supported, how much influence will the model have, and what is the consequence of error?	Credibility requirements should be proportionate to the intended question and risk.	Converts evidence into a bounded statement of permissible use.	Local evidence may be generalized to other decisions, populations, or regulatory contexts.	Qualification for one use does not confer universal validity, acceptance, or deployment readiness.

Figure 1 presents the mechanistic accountability hybrid model, showing how the article's key components and boundaries are connected within why hybrid models need mechanistic accountability. The central anchor is not a computational pipeline that automatically converts data into a qualified decision. It is an accountability structure in which learned inputs, PBPK states, and predictions remain connected by explicit semantic and evidentiary obligations. Identifiability, uncertainty propagation,

biological plausibility, verification, validation, and simpler-model comparison operate as complementary restrictions on interpretation. Their function is not to certify a hybrid model universally, but to determine which claims remain supportable in a defined context and which claims must be weakened, deferred, or rejected.

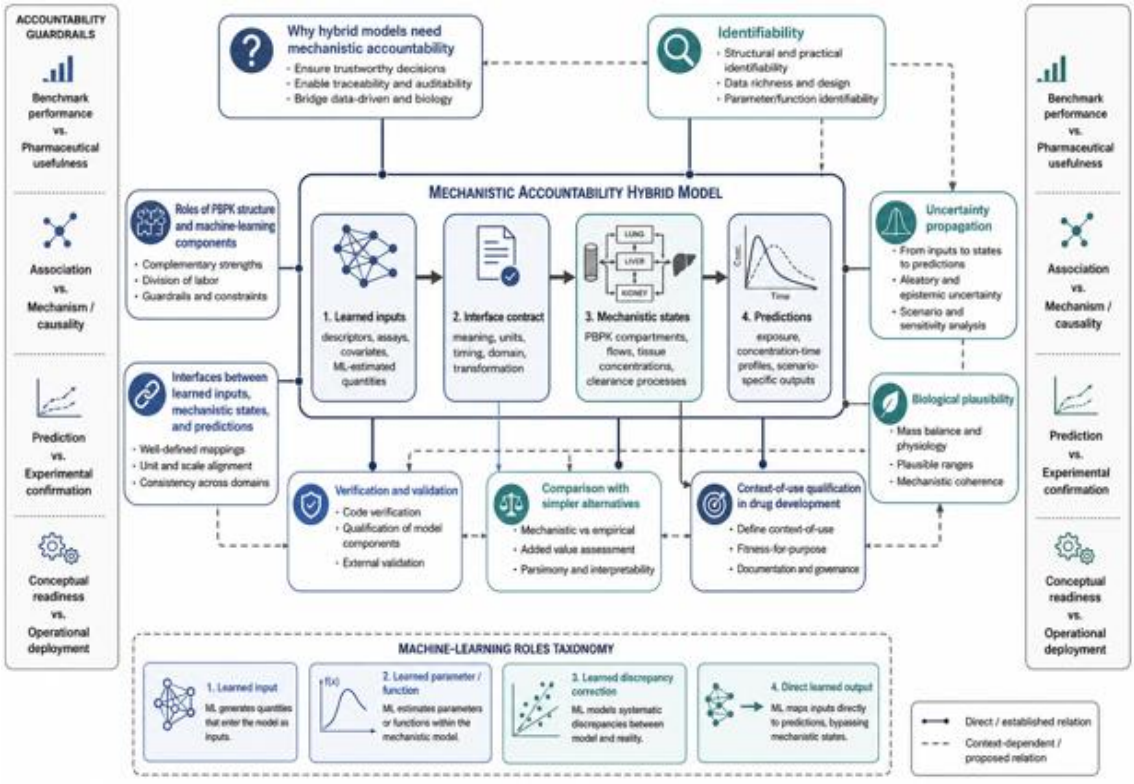

Figure 1. Mechanistic Accountability Hybrid Model.

The figure is an original conceptual synthesis that organizes the article's central contribution across hybrid models need mechanistic accountability, roles of PBPK structure and machine-learning components, interfaces between learned inputs, mechanistic states, and predictions, interfaces between learned inputs, mechanistic states, identifiability, uncertainty propagation, and biological plausibility. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Roles of PBPK Structure and Machine-Learning Components

PBPK structure should be understood as a scientifically useful representation of physiological organization rather than as a literal reproduction of biological reality. Even widely used clearance formulations depend on simplifying assumptions concerning organ mixing, blood flow, binding, permeability, and the relationship between intrinsic and observed clearance [10]. This distinction is central to mechanistic accountability. A hybrid model cannot be justified merely by placing a learned component inside an established PBPK scaffold, because the scaffold itself contains abstractions, uncertain parameters, and context-dependent simplifications. The PBPK contribution is therefore not automatic truth but structured constraints: explicit compartments, mass-balance relations, physiological flows, interpretable state variables, and a framework for simulating changes across regimens, populations, and interacting processes. These properties can strengthen scientific reasoning only when their assumptions and limitations remain visible.

Machine learning can contribute to PBPK modeling by estimating ADME-related inputs, learning complex covariate relationships, approximating difficult submodels, identifying patterns in high-dimensional evidence, or serving as a computational surrogate [11]. Each role carries a different interpretation. An ML-predicted partition coefficient or intrinsic clearance is an upstream estimate used by the mechanistic model. A learned covariate function modifies how a parameter varies across individuals or conditions. A surrogate model approximates the outputs of a computationally demanding simulator. A residual-correction model estimates discrepancies that the mechanistic equations do not explain. A direct concentration–time predictor may bypass mechanistic state estimation almost entirely. Mechanistic accountability therefore requires the machine-learning component to be named by function rather than described generically as an "AI layer." The role determines what evidence is relevant, what uncertainty must be propagated, and what mechanistic claims remain permissible.

Comparisons between experimentally derived and machine-learned PBPK inputs illustrate why role definition is insufficient without evidence about the quantity being transferred. Learned compound properties may support PBPK prediction when direct measurements are unavailable, but their usefulness remains conditional on the chemical domain, endpoint definition,

assay quality, training data, and availability of credible alternatives [12]. A prediction generated from molecular structure is not interchangeable with a value measured under a specified experimental protocol. Even when both enter the same PBPK equation, they carry different uncertainties and may reflect different operational definitions. Mechanistic accountability therefore requires a distinction among measured inputs, literature-derived priors, fitted parameters, learned estimates, and latent corrective terms. Their numerical values may occupy the same field in software, but their epistemic status is not the same.

Hybrid architectures can also divide labor by allowing machine learning to estimate parameters or nonlinear functions while retaining PBPK equations for state evolution and scenario simulation [13]. By contrast, a neural model that predicts concentration–time profiles directly performs a different representational task and should not be described as mechanistically explanatory merely because its outputs resemble those of a pharmacokinetic model [14]. The proposed construct consequently distinguishes four principal roles: learned input, learned parameter or function, learned discrepancy correction, and direct learned output. The first three may interact with mechanistic states, whereas the fourth may generate predictions without an explicit physiological state interpretation. None is intrinsically preferable. Their acceptability depends on the intended use, data support, interface transparency, uncertainty treatment, validation design, and whether the added learning component produces a defensible advantage over simpler alternatives.

Interfaces between Learned Inputs, Mechanistic States, and Predictions

The scientific meaning of a hybrid model is determined not only by its components but by the interfaces through which information passes between them. A learned quantity may enter before simulation as a compound-specific input, within simulation as a parameter or time-varying function, after simulation as a discrepancy correction, or outside the mechanistic structure as a direct prediction. Neural differential-equation approaches demonstrate that learned dynamics can be coupled to pharmacokinetic structure, making the position of the learned component an explicit architectural decision rather than an incidental implementation detail [15]. Mechanistic accountability therefore requires an interface contract that specifies the quantity's definition, unit, transformation, timing, directionality, allowable range, domain of applicability, and uncertainty before it can influence a PBPK state.

This contract is necessary because machine-learning and pharmacometric models can produce superficially similar outputs while encoding different assumptions about time, variability, covariate effects, and extrapolation. Comparative pharmacokinetic modeling shows that these approaches differ in their inductive biases and in how they behave when asked to predict beyond the conditions represented in the available observations [16]. A concentration–time curve alone cannot reveal whether the model conserved mass, separated interindividual from residual variability, represented a physiologically meaningful compartment, or learned a dataset-specific association. Outputs should therefore not be exchanged between learned and mechanistic components without documenting what information is preserved, transformed, or lost at the interface.

The interface becomes particularly consequential when machine learning translates upstream chemical or formulation information into parameters that determine PBPK trajectories. AI-assisted PBPK applications illustrate how learned mappings from physicochemical or formulation-related attributes to model parameters can become upstream determinants of tissue distribution and exposure predictions [17]. In such cases, the PBPK equations may propagate a learned estimate consistently while also amplifying its error across tissues, time points, or simulated populations. The physiological structure can therefore make the consequences of an input estimate more interpretable, but it cannot establish that the estimate is correct. Accountability requires the uncertainty, applicability domain, and evidentiary origin of the learned quantity to accompany it through the mechanistic model.

Existing implementations use machine learning for distinct purposes, including covariate identification, parameter estimation, residual modeling, and direct prediction [18]. These roles should not be collapsed into a single category because they create different failure modes and validation requirements. Physiological constraints, parameter-specific learning branches, and bounded functional forms may reduce implausible behavior and make learned covariate effects more inspectable [19]. Nevertheless, a constraint does not prove that a learned effect is causal or biologically complete. The proposed interface framework therefore asks three questions at every boundary: what is being transferred, how does it alter mechanistic states, and which downstream claims remain defensible after its uncertainty and assumptions have been considered.

Identifiability, Uncertainty Propagation, and Biological Plausibility

Identifiability is a prerequisite for interpreting internal hybrid-model quantities as mechanistically informative. Structural identifiability concerns whether a model's equations and observation structure permit unique parameter recovery in principle, whereas practical identifiability concerns whether the available data constrain those parameters sufficiently in practice [20]. Hybridization can affect both. A flexible learned function may create new parameter symmetries, absorb information previously assigned to a physiological parameter, or reproduce the same observations through several internal configurations. A model may consequently generate stable predictions while leaving the mechanistic interpretation of its parameters unresolved.

This distinction is important because apparently adequate output behavior can coexist with non-identifiable parameterization. In quantitative pharmacological models, insufficiently constrained parameters weaken the interpretation of mechanisms and reduce confidence in predictions made under new interventions or conditions [21]. For a hybrid PBPK model, the relevant unit of analysis is not only the PBPK parameter or machine-learning weight considered separately. It is the combined mapping

from data to learned quantity, from learned quantity to mechanistic state, and from state to observation. Mechanistic accountability therefore requires claims to be restricted to quantities that are structurally distinguishable and sufficiently informed by the available evidence.

Uncertainty should be represented as an end-to-end property of this mapping rather than as an isolated confidence value reported by the machine-learning component. Parameter-estimation practice in systems modeling emphasizes that uncertainty in model inputs and parameters propagates into predicted behavior [22]. Hybrid systems add uncertainty from training labels, model selection, applicability-domain limitations, learned transformations, PBPK structural assumptions, population physiology, and residual discrepancy. These sources may be correlated and may affect different scenarios unequally. A narrow prediction interval produced by only one component can therefore be misleading. The proposed uncertainty propagation ledger records each material source, its dependency structure, the state variables it affects, and its contribution to the final prediction and decision boundary.

Biological plausibility provides a complementary test but must not be treated as proof of mechanistic correctness. Mechanistic prediction depends on the relationship among model structure, parameter estimation, available data, and the claim being made [23]. Parameter ranges, mass-balance behavior, directional responses, organ-specific trajectories, and responses to known perturbations can reveal whether acceptable predictions arise through scientifically credible behavior. Current model-assessment practice is multidimensional and varies with modeling purpose, reinforcing the need to examine biological behavior alongside numerical performance [24]. A plausible model may still be non-identifiable or predictively inadequate, while an accurate model may achieve agreement through compensating errors. Accountability therefore requires identifiability, propagated uncertainty, and biological challenge testing to constrain one another.

Verification, Validation, and Comparison with Simpler Alternatives

Credibility assessment for mechanistic pharmaceutical models requires transparent reporting, justified inputs, verification, evaluation, and a risk-informed relationship between the model's context of use, influence, and consequences of error [25–27]. For hybrid PBPK–machine-learning models, these requirements must apply to the complete coupled system as well as to each component. Verification should establish that the intended equations, software operations, transformations, units, constraints, and interface rules have been implemented correctly. Validation should then examine whether the model produces sufficiently credible behavior for the specified scientific question. Neither activity can be replaced by reporting a favorable prediction metric.

Component-level verification should include machine-learning data processing, parameter scaling, software versioning, unit conversion, interface timing, solver behavior, and reproducibility under controlled test cases. PBPK-specific checks may include conservation relations, limiting cases, parameter bounds, and independent calculations for selected scenarios. Machine-learning-specific checks may include label integrity, leakage detection, deterministic preprocessing, and consistency between training and inference pipelines. Passing these tests establishes implementation fidelity, not biological validity. A correctly coded model may still represent an unsuitable endpoint, contain structurally incorrect assumptions, or extrapolate beyond the evidence supporting its learned component.

Validation intensity should be proportionate to the importance of the model output and the cost and risk associated with additional evidence rather than imposed through a universal checklist [28]. A model used for exploratory compound prioritization may be evaluated through retrospective external data, sensitivity analysis, and clearly stated uncertainty boundaries. A model intended to materially influence a high-consequence development decision may require stronger external evidence, prospective challenge cases, independent review, and predefined failure criteria. The proposed accountability construct therefore treats validation as a structured argument: the evidence should show why the model is adequate for a particular purpose while also identifying which uncertainties and unsupported scenarios remain.

Comparison with simpler alternatives is essential because increased architectural complexity does not itself establish scientific value. Model assessment should examine whether biologically relevant behaviors and relationships are reproduced rather than focusing only on selected output agreement [29]. A hybrid model should therefore be compared under matched conditions with an appropriate conventional PBPK model, empirical pharmacokinetic model, or purely data-driven predictor. The comparison should consider calibration, robustness, extrapolation, biological behavior, uncertainty, interpretability, and decision relevance. A small improvement in one predictive metric may not justify a substantially less identifiable or less auditable model. Conversely, a hybrid model may be warranted when it produces reproducible, context-relevant gains that simpler alternatives cannot achieve without stronger assumptions or loss of necessary physiological structure.

Context-Of-Use Qualification in Drug Development

Context of use defines the specific question a model addresses, the conditions under which it will be applied, the influence assigned to its output, and the consequence of an incorrect conclusion. Early regulatory engagement can help clarify the drug-development question and the evidence expected from a modeling analysis [30]. For a hybrid PBPK–machine-learning model, the context statement should identify the compounds, formulations, populations, physiological conditions, dosing regimens, interaction mechanisms, and prediction endpoints within scope. It should also state whether the model is intended for hypothesis generation, experimental prioritization, scenario exploration, evidence integration, or material decision support.

Alignment among model developers, pharmaceutical scientists, decision-makers, and regulatory stakeholders can reduce ambiguity about the analysis plan and the permissible interpretation of model outputs [31]. Mechanistic accountability adds

component-level detail to this alignment. The context statement should identify which inputs are measured, learned, fitted, or assumed; which PBPK states are considered mechanistically interpretable; which uncertainties are propagated; and which failure conditions trigger reduced model influence or additional evidence generation. Human oversight should be assigned explicitly rather than invoked generically. Data scientists, DMPK scientists, pharmacometricians, clinicians, and decision owners may each be responsible for different parts of the accountability chain.

Efforts toward harmonized model-informed drug-development principles emphasize clear questions, fit-for-purpose evidence, transparent communication, and consistent terminology [32]. These principles support a tiered approach in which evidence depth increases with model influence and consequence of error. The proposed framework does not assign universal categories or numerical acceptance thresholds. Instead, it asks whether verification, identifiability, uncertainty, biological plausibility, predictive evaluation, and comparator evidence are collectively sufficient for the proposed use. A model may therefore be credible for one compound class, population, or scenario while remaining unsuitable for another, even when the underlying software and architecture are unchanged.

Contemporary regulatory use of PBPK analysis indicates that acceptance depends on adequate qualification for the submitted question and on the evidence supporting the model's proposed role [33]. Formal model-informed development meeting pathways may facilitate iterative discussion of purpose, evidence, and decision influence, but they do not provide automatic acceptance of a particular architecture [34]. Context-of-use qualification should therefore conclude with a bounded statement specifying what the model may support, what it cannot establish, and what additional evidence would be required to expand its influence. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop context-of-use qualification in drug development within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Mechanistic Accountability in Hybrid Physiologically Based Pharmacokinetic and Machine-Learning Models for Drug

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Drug-development question	Proposed decision, population, regimen, endpoint, and consequence of error	Defines why the model is being developed and how its output may be used	Explicit context-of-use statement	Questions may be framed too broadly or altered after model evaluation	Prospective agreement on scope, model influence, and failure consequences
Evidence and data layer	Chemical descriptors, in vitro ADME, formulation data, preclinical observations, clinical data, and population covariates	Supplies information for both learned and mechanistic components	Versioned, provenance-linked analysis dataset	Missingness, measurement error, assay heterogeneity, selection bias, and distribution shift	Data-quality assessment and demonstration of suitability for the intended use
Learned input layer	Labeled training data and relevant descriptors or covariates	Estimates a PBPK input, parameter function, latent representation, or residual term	Learned quantity with defined unit or interpretation	Prediction error, domain limitations, proxy labels, and unstable associations	External evaluation, calibration, applicability-domain analysis, and sensitivity testing
Interface contract	Learned output, transformation rules, timing, units, and constraints	Transfers information from the ML component to the PBPK model without semantic ambiguity	Auditable mapping into a specified parameter, function, state, or correction	Unit errors, hidden transformations, timing mismatch, or inconsistent definitions	Dimensional analysis, boundary tests, independent implementation review, and reproducibility checks
PBPK structural layer	Physiological parameters, drug-specific inputs, population characteristics, and learned quantities	Simulates state evolution through physiologically organized equations	Tissue amounts, concentrations, clearances, or related mechanistic states	Structural simplification, uncertain physiology, parameter correlation, and misspecification	Verification, sensitivity analysis, identifiability assessment, and biological challenge testing
Observation and prediction layer	Mechanistic states, observation model, residual structure, and scenario definition	Translates internal states into reportable exposure predictions	Prediction for a defined compound, population, regimen, and endpoint	Measurement-model error, unsupported extrapolation, and scenario dependence	External predictive evaluation matched to the proposed application
Uncertainty propagation layer	Data uncertainty, ML uncertainty, parameter uncertainty, structural uncertainty, and model discrepancy	Carries uncertainty through interfaces, states, and predictions	Scenario-specific predictive distributions and uncertainty attribution	Dependencies may be unknown or omitted; estimates remain conditional on model assumptions	Calibration assessment, probabilistic sensitivity analysis, and empirical coverage evaluation
Biological accountability layer	Physiological ranges, conservation principles, known directional responses, and perturbation evidence	Tests whether internal behavior remains scientifically credible	Plausibility assessment and identified failure conditions	Plausibility may be incomplete and does not prove causal correctness	Predefined challenge scenarios and expert review linked to the intended use
Comparative adequacy layer	Hybrid model and matched simpler PBPK, empirical, or ML alternatives	Determines whether hybrid complexity provides decision-relevant benefit	Comparative evidence on calibration, robustness,	Comparator choice may bias conclusions; no single metric captures total value	Locked comparison protocol using common data, scenarios, and endpoints

			extrapolation, and interpretability		
Human oversight and governance layer	Complete model dossier, uncertainty summary, validation evidence, and decision context	Assigns responsibility for scientific review, use restrictions, and escalation	Traceable approval, rejection, or limited-use rationale within the organization	Diffuse responsibility or overreliance on software outputs	Independent multidisciplinary review and documented decision ownership
Context-of-use qualification layer	Integrated evidence from all preceding components	Determines the permissible influence of the hybrid model	Bounded qualification statement with limitations and contingencies	Qualification may be generalized beyond the evaluated use	Fit-for-purpose credibility assessment and requalification after material model or context changes
Research and implementation layer	Identified evidence gaps, failed challenges, user feedback, and monitoring findings	Guides staged improvement without presuming readiness	Prioritized studies, reporting improvements, and infrastructure needs	Research priorities may be mistaken for completed validation	Prospective evaluation before expanding claims, populations, or decision influence

Limitations and Boundary Conditions

The proposed construct is conceptual and has not been empirically validated as a complete framework. The selected literature supports its constituent concerns—model credibility, PBPK qualification, machine-learning limitations, interface design, identifiability, uncertainty, and context dependence—but does not demonstrate that the proposed organization improves development decisions. Future work must determine whether mechanistic accountability can be applied reproducibly across modeling teams, software platforms, therapeutic areas, and stages of development. It must also establish whether the additional documentation and evaluation burden produces benefits proportionate to its cost.

The construct cannot guarantee that a hybrid model represents the correct biological mechanism. Identifiability analysis may expose ambiguity but cannot remove uncertainty unsupported by data. Biological plausibility testing can reject some implausible behaviors but cannot prove causal completeness. Uncertainty propagation remains conditional on the sources and dependencies represented in the analysis. External validation may establish predictive adequacy within an evaluated domain but cannot ensure performance under all future compounds, populations, formulations, or physiological states. Mechanistic accountability should therefore constrain claims rather than serve as a certificate of model truth.

The framework also does not establish clinical utility, regulatory acceptability, therapeutic value, or deployment readiness. It should not be used to convert exploratory predictions into dose recommendations, replace necessary experiments without sufficient evidence, or imply that PBPK structure legitimizes an otherwise unsupported learned relationship. Context-of-use boundaries may change when the model architecture, training data, software implementation, population, regimen, or decision consequence changes. Reuse should therefore trigger reassessment rather than automatic transfer of prior credibility. The principal misuse risk is rhetorical: a model may be labeled accountable, mechanistic, or interpretable without supplying the evidence required by those terms.

Research Agenda and Implementation Priorities

The first research priority is to make hybrid architecture and interface claims testable. Studies should compare clearly defined interface classes—learned inputs, learned parameter functions, learned discrepancy corrections, and direct learned outputs—under matched pharmacokinetic questions. Evaluation should determine how interface placement affects identifiability, uncertainty calibration, biological behavior, robustness, and extrapolation. Hybrid-specific identifiability methods are needed for flexible functions and latent representations that interact with conventional PBPK parameters. Prospective experiments should also investigate whether uncertainty from learned inputs is adequately propagated and whether its decomposition remains stable under distribution shift.

A second priority is the development of shared reporting and evaluation infrastructure. Hybrid-model reports should include data provenance, endpoint definitions, interface contracts, component-specific assumptions, software verification, applicability domains, uncertainty dependencies, biological challenge tests, and simpler-model comparisons. Open benchmark cases should contain known failure modes rather than merely rewarding predictive accuracy. Such cases could test unit inconsistency, parameter confounding, hidden extrapolation, compensating errors, implausible tissue trajectories, and misleadingly narrow uncertainty. Benchmark outcomes should remain methodological and should not be presented as substitutes for application-specific qualification.

Implementation should proceed in stages. Initial use may focus on low-influence activities such as hypothesis generation, experiment prioritization, and identification of uncertainty-sensitive parameters. Expansion to more influential development applications should depend on prospective evidence, independent review, predefined failure criteria, and clear reassessment triggers. Organizational responsibilities should be distributed across data science, PBPK modeling, DMPK, clinical pharmacology, quality assurance, and decision ownership. The objective is not to create an administrative checklist but to ensure that every learned contribution to a mechanistic prediction has a visible scientific justification and a bounded role.

Conclusion

Mechanistic accountability provides a proposed organizing principle for evaluating hybrid PBPK–machine-learning models without treating physiological structure, explanatory displays, or predictive performance as sufficient evidence of mechanism or pharmaceutical usefulness. It links learned quantities to their provenance, semantic interfaces, effects on mechanistic states, identifiability, propagated uncertainty, biological plausibility, verification, validation, comparator evidence, and context-specific decision influence. The framework does not establish that hybrid models are universally preferable, empirically validated, regulator-approved, clinically useful, or ready for deployment. Its contribution is instead to define the conditions under which a hybrid prediction may support a bounded scientific claim and the conditions under which interpretation must remain restricted. By making component roles and evidentiary boundaries explicit, mechanistic accountability may support more transparent theory building, evaluation, and communication in model-informed drug development.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Musuamba FT, Skottheim Rusten I, Lesage R, Russo G, Bursi R, Emili L, et al. Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. CPT Pharmacometrics Syst Pharmacol. 2021;10(8):804-25. doi:10.1002/psp4.12669
2. Luzon E, Blake K, Cole S, Nordmark A, Versantvoort C, Berglund EG. Physiologically based pharmacokinetic modeling in regulatory decision-making at the European Medicines Agency. Clin Pharmacol Ther. 2017;102(1):98-105. doi:10.1002/cpt.539
3. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
4. Wang Y, Zhu H, Madabushi R, Liu Q, Huang SM, Zineh I. Model-informed drug development: Current US regulatory practice and future considerations. Clin Pharmacol Ther. 2019;105(4):899-911. doi:10.1002/cpt.1363
5. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15. doi:10.1038/s42256-019-0048-x
6. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11). doi:10.1016/S2589-7500(21)00208-9
7. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284-6. doi:10.1126/science.abg1834
8. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
9. McComb M, Bies R, Ramanathan M. Machine learning in pharmacometrics: Opportunities and challenges. Br J Clin Pharmacol. 2022;88(4):1482-99. doi:10.1111/bcp.14801
10. Pang KS, Han YR, Noh K, Lee PI, Rowland M. Hepatic clearance concepts and misconceptions: Why the well-stirred model is still used even though it is not physiologic reality? Biochem Pharmacol. 2019;169:113596. doi:10.1016/j.bcp.2019.07.025
11. Chou WC, Lin Z. Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. Toxicol Sci. 2023;191(1):1-14. doi:10.1093/toxsci/kfac101
12. Parrott N, Manevski N, Olivares-Morales A. Can we predict clinical pharmacokinetics of highly lipophilic compounds by integration of machine learning or in vitro data into physiologically based models? A feasibility study based on 12 development compounds. Mol Pharm. 2022;19(11):3858-68. doi:10.1021/acs.molpharmaceut.2c00350
13. Li Y, Wang Z, Li Y, Du J, Gao X, Li Y, et al. A combination of machine learning and PBPK modeling approach for pharmacokinetics prediction of small molecules in humans. Pharm Res. 2024;41(7):1369-79. doi:10.1007/s11095-024-03725-y
14. Bräm DS, Parrott N, Hutchinson L, Steiert B. Introduction of an artificial neural network-based method for concentration-time predictions. CPT Pharmacometrics Syst Pharmacol. 2022;11(6):745-54. doi:10.1002/psp4.12786
15. Lu J, Deng K, Zhang X, Liu G, Guan Y. Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens. iScience. 2021;24(7):102804. doi:10.1016/j.isci.2021.102804
16. Keutzer L, You H, Farnoud A, Nyberg J, Wicha SG, Maher-Edwards G, et al. Machine learning and pharmacometrics for prediction of pharmacokinetic data: Differences, similarities and challenges illustrated with rifampicin. Pharmaceutics. 2022;14(8):1530. doi:10.3390/pharmaceutics14081530

17. Chou WC, Chen Q, Yuan L, Cheng YH, He C, Monteiro-Riviere NA, et al. An artificial intelligence-assisted physiologically-based pharmacokinetic model to predict nanoparticle delivery to tumors in mice. J Control Release. 2023;361:53-63. doi:10.1016/j.jconrel.2023.07.040
18. Janssen A, Bennis FC, Mathôt RAA. Adoption of machine learning in pharmacometrics: An overview of recent implementations and their considerations. Pharmaceutics. 2022;14(9):1814. doi:10.3390/pharmaceutics14091814
19. Janssen A, Bennis FC, OPTI-CLOT Study Group and SYMPHONY Consortium, Cnossen MH, Mathôt RAA. On inductive biases for the robust and interpretable prediction of drug concentrations using deep compartment models. J Pharmacokinet Pharmacodyn. 2024;51(4):355-66. doi:10.1007/s10928-024-09906-x
20. Wieland FG, Hauber AL, Rosenblatt M, Tönsing C, Timmer J. On structural and practical identifiability. Curr Opin Syst Biol. 2021;25:60-9. doi:10.1016/j.coisb.2021.03.005
21. Sher A, Niederer SA, Mirams GR, Kirpichnikova A, Allen R, Pathmanathan P, et al. A quantitative systems pharmacology perspective on the importance of parameter identifiability. Bull Math Biol. 2022;84(3):39. doi:10.1007/s11538-021-00982-5
22. Mitra ED, Hlavacek WS. Parameter estimation and uncertainty quantification for systems biology models. Curr Opin Syst Biol. 2019;18:9-18. doi:10.1016/j.coisb.2019.10.006
23. Fröhlich F, Kessler T, Weindl D, Shadrin A, Schmiester L, Hache H, et al. Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. Cell Syst. 2018;7(6):567-79. doi:10.1016/j.cels.2018.10.013
24. Chan JR, Allen R, Boras B, Cabal A, Damian V, Gibbons FD, et al. Current practices for QSP model assessment: An IQ consortium survey. J Pharmacokinet Pharmacodyn. 2024;51(5):543-55. doi:10.1007/s10928-022-09811-1
25. Shebley M, Sandhu P, Emami Riedmaier A, Jamei M, Narayanan R, Patel A, et al. Physiologically based pharmacokinetic model qualification and reporting procedures for regulatory submissions: A consortium perspective. Clin Pharmacol Ther. 2018;104(1):88-110. doi:10.1002/cpt.1013
26. Peters SA, Dolgos H. Requirements to establishing confidence in physiologically based pharmacokinetic models and overcoming some of the challenges to meeting them. Clin Pharmacokinet. 2019;58(11):1355-71. doi:10.1007/s40262-019-00790-0
27. Kuemmel C, Yang Y, Zhang X, Florian J, Zhu H, Tegenge M, et al. Consideration of a credibility assessment framework in model-informed drug development: Potential application to physiologically based pharmacokinetic modeling and simulation. CPT Pharmacometrics Syst Pharmacol. 2020;9(1):21-8. doi:10.1002/psp4.12479
28. Lemaire V, Hu C, van der Graaf PH, Wang W. No recipe for quantitative systems pharmacology model validation, but a balancing act between risk and cost. Clin Pharmacol Ther. 2024;115(1):25-8. doi:10.1002/cpt.3082
29. Singh FA, Afzal N, Smithline SJ, Thalhauser CJ. Assessing the performance of QSP models: Biology as the driver for validation. J Pharmacokinet Pharmacodyn. 2024;51(5):533-42. doi:10.1007/s10928-023-09871-x
30. Madabushi R, Benjamin JM, Grewal R, Pacanowski MA, Strauss DG, Wang Y, et al. The US Food and Drug Administration's model-informed drug development paired meeting pilot program: Early experience and impact. Clin Pharmacol Ther. 2019;106(1):74-8. doi:10.1002/cpt.1457
31. Galluppi GR, Brar S, Caro L, Chen Y, Frey N, Grimm HP, et al. Industrial perspective on the benefits realized from the FDA's model-informed drug development paired meeting pilot program. Clin Pharmacol Ther. 2021;110(5):1172-5. doi:10.1002/cpt.2265
32. Marshall S, Ahamadi M, Chien J, Iwata D, Farkas P, Filipe A, et al. Model-informed drug development: Steps toward harmonized guidance. Clin Pharmacol Ther. 2023;114(5):954-9. doi:10.1002/cpt.3006
33. Paul P, Colin PJ, Musuamba Tshinanu F, Versantvoort C, Manolis E, Blake K. Current use of physiologically based pharmacokinetic modeling in new medicinal product approvals at EMA. Clin Pharmacol Ther. 2025;117(3):808-17. doi:10.1002/cpt.3525
34. Madabushi R, Benjamin J, Zhu H, Zineh I. The US Food and Drug Administration's model-informed drug development meeting program: From pilot to pathway. Clin Pharmacol Ther. 2024;116(2):278-81. doi:10.1002/cpt.3228