Pharmacophore, 15(5) 2024, Pages: 90-99

Pharmacophore

ISSN-2229-5402

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

WHEN IS AN ARTIFICIAL INTELLIGENCE MODEL READY TO INFLUENCE DECISIONS IN EARLY-STAGE PHARMACEUTICAL DISCOVERY?

Sarah Jensen[1]*, Pieter Vos[2], Lars Olsen[3]

1. ***Department of AI Readiness and Decision Science, Faculty of Pharmaceutical Sciences, University of Copenhagen, Copenhagen, Denmark.***
2. ***Department of Model Validation and Translational Informatics, Faculty of Pharmaceutical Sciences, Wageningen University, Wageningen, Netherlands.***
3. ***Department of Early-Stage Drug Discovery and Computational Decision Support, Faculty of Pharmacy, University of Queensland, Brisbane, Australia.***

ARTICLE INFO

Received:
01 June 2024
Received in revised form:
11 October 2024
Accepted:
16 October 2024
Available online:
28 October 2024

Keywords: Artificial intelligence, Early drug discovery, Decision readiness, Evidence adequacy, Uncertainty quantification, Transportability

ABSTRACT

Artificial intelligence models increasingly contribute to target assessment, molecular screening, property prediction, candidate generation, and evidence integration in early pharmaceutical discovery. However, the ability to produce accurate retrospective predictions does not establish that a model is ready to influence a consequential scientific or portfolio decision. Existing evaluation practices frequently emphasize benchmark performance while leaving the intended decision, consequences of error, data suitability, transportability, uncertainty, experimental corroboration, and human accountability insufficiently specified. This Original Decision-Readiness Model Article addresses that unresolved problem by proposing a decision-centred approach to evaluating artificial intelligence models before their outputs are permitted to shape early-discovery choices. The proposed model treats readiness as a conditional relationship among four interdependent components: decision context, consequence and risk, evidence adequacy, and human and organizational oversight. It further distinguishes levels of permissible influence, ranging from exploratory analysis to experimentally anchored and portfolio-influencing use. Under this approach, readiness cannot be inferred from a single performance measure or transferred automatically between targets, assays, chemical spaces, laboratories, projects, or decision stages. Instead, the evidentiary burden must increase with the consequences and irreversibility of the proposed use. Validation, calibrated uncertainty, transportability analysis, failure-mode testing, and explicit decision governance operate as cross-cutting requirements. The model is a conceptual contribution rather than an empirically validated maturity standard, regulatory instrument, or deployment protocol. Its purpose is to make claims about artificial intelligence readiness more precise, contestable, and proportionate to pharmaceutical decision risk. Future work must operationalize its components, test the reproducibility of readiness assignments, and determine whether decision-centred evaluation improves prospective scientific and portfolio outcomes.

To Cite This Article: Jensen S, Vos P, Olsen L. When Is an Artificial Intelligence Model Ready to Influence Decisions in Early-Stage Pharmaceutical Discovery? Pharmacophore. 2024;15(5):90-9. https://doi.org/10.51847/EXrfjyPbrw

Introduction

Artificial intelligence now supports target identification, molecular-property prediction, compound prioritization, de novo design, and increasingly automated discovery workflows, yet broad technical capability becomes pharmaceutically meaningful only when it is tied to a defined discovery problem and, where appropriate, to experimental feedback [1-3]. This distinction is important because early pharmaceutical discovery is not a single prediction task. It is a sequence of heterogeneous decisions about whether a biological hypothesis is sufficiently credible, whether a chemical series deserves additional investigation, which compounds should be tested, and when limited resources should be redirected or committed. A model that performs well in one analytical setting may therefore be useful for exploratory investigation while remaining insufficiently supported for a decision that changes experimental priorities or project funding.

The unresolved question is not whether artificial intelligence can generate predictions, rankings, explanations, or molecular proposals. It is when those outputs have acquired enough evidentiary and organizational support to influence a particular

Corresponding Author: Sarah Jensen; Department of AI Readiness and Decision Science, Faculty of Pharmaceutical Sciences, University of Copenhagen, Copenhagen, Denmark. E-mail: sarah.jensen@food.ku.dk.

pharmaceutical decision. Critical analyses of artificial intelligence in drug discovery have emphasized that realistic impact depends on problem formulation, the relevance and quality of available evidence, the way performance is evaluated, and the downstream actions attached to model outputs [2]. Nevertheless, readiness is still often discussed through model-centred language: accuracy, discrimination, enrichment, optimization, novelty, or computational efficiency. These properties may be necessary for some uses, but they do not define whether a model is suitable for the decision, whether errors are tolerable, whether uncertainty is actionable, or whether the output is integrated into a scientifically accountable workflow.

This gap has practical and epistemic consequences. In early discovery, a prediction may contribute to hypothesis generation, candidate exclusion, assay prioritization, synthetic planning, or project progression. Each use creates a different burden of evidence because the consequences, alternatives, reversibility, and opportunities for experimental correction differ. A model used to explore an unfamiliar chemical space does not require the same justification as a model used to deprioritize a target or advance a compound series. Similarly, a model connected to an iterative design–make–test cycle occupies a different evidentiary position from an isolated model evaluated only through retrospective data [3]. Treating these cases as equivalent obscures the difference between computational competence and warranted decision influence.

This article proposes a pharmaceutical decision-readiness model organized around four interdependent components: context, risk, evidence, and oversight. The model reframes readiness as a bounded property of a model–decision relationship rather than an intrinsic property of an algorithm. It also proposes readiness levels that restrict the degree of influence a model may exert, from exploratory analysis to portfolio-influencing use. The intended contribution is conceptual and methodological: to provide a structured basis for asking what evidence is required before a model output may legitimately alter early-discovery action. The model does not claim empirical validation, universal applicability, regulatory acceptance, or operational deployment readiness. Instead, it identifies relationships and boundaries that require prospective evaluation in target selection, screening, molecular design, and lead progression.

Decision Points in Early Pharmaceutical Discovery and their Evidentiary Demands

Early pharmaceutical discovery involves decision points whose objects, evidence requirements, and failure consequences differ substantially. Target assessment, for example, requires more than a prediction that a gene, protein, or pathway is associated with disease. A credible target case may require evidence concerning disease linkage, biological mechanism, intervention feasibility, safety liabilities, translational relevance, and the availability of experiments capable of resolving remaining uncertainty [4]. These domains do not contribute interchangeable information. Strong evidence in one domain cannot automatically compensate for a critical deficiency in another. A target with compelling disease association may remain unsuitable when its mechanism is poorly understood, modulation is technically infeasible, or plausible safety risks cannot be bounded. Decision readiness must therefore be assessed against the evidentiary structure of the particular target decision rather than against the apparent strength of an isolated model output.

Human genetics illustrates both the value and limitations of a high-priority evidence domain. Genetic evidence can strengthen the rationale for prioritizing a target by connecting naturally occurring variation to disease-relevant phenotypes and by helping to distinguish targets with different levels of human support [5]. However, genetic support does not independently establish that pharmacological modulation will reproduce the relevant biological effect, that the direction and magnitude of modulation are therapeutically achievable, or that the target will be safe and chemically tractable. In the proposed decision-readiness model, genetic evidence is therefore neither a decorative input nor a deterministic approval criterion. Its role depends on the target hypothesis, disease mechanism, intervention modality, supporting biological evidence, and the specific decision under consideration.

Evidence associated with improved development prospects must likewise be interpreted probabilistically rather than as a guarantee of success. Modelling work examining human-genetic support and drug-development outcomes suggests that genetically supported mechanisms may have better development prospects, while also showing that the magnitude and interpretation of this association depend on target–indication definitions, data construction, and analytical assumptions [6]. For decision readiness, the implication is not that a model should automatically favour genetically supported targets. Rather, the model should expose how genetic evidence changes the comparative case, what uncertainties remain, and whether the decision would be materially different under alternative evidence definitions. Association with downstream success cannot substitute for mechanism, causal confirmation, tractability, or project-specific experimental evidence.

Target prioritization and related early-discovery choices commonly require the integration of heterogeneous sources, including genetics, functional genomics, molecular biology, disease knowledge, existing pharmacology, safety information, and chemical tractability. Evidence-integration resources demonstrate the scientific value of organizing these materials into traceable target–disease relationships rather than treating any single source as determinative [7]. The proposed model extends this integration logic by requiring decision teams to specify the decision, identify the evidence domains that could change it, distinguish observed evidence from model-derived inference, and document unresolved contradictions. Data availability is not equivalent to evidentiary completeness, and an integrated score is not equivalent to a justified decision. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop decision points in early pharmaceutical discovery and their evidentiary demands within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Decision points in early pharmaceutical discovery and their evidentiary demands

Construct or Evidence Domain	Core Question	Scientific Basis	Role in the Article	Failure or Overclaiming Risk	Boundary Statement
Decision Specification	What exact choice may the model influence, for whom, at what discovery stage, and relative to which alternatives?	A model output has meaning only in relation to a defined action, endpoint, user, and time horizon.	Establishes the unit of decision readiness and prevents generic claims about model capability.	Evaluating a convenient prediction task rather than the actual pharmaceutical decision.	Readiness for one decision does not imply readiness for another decision involving a different endpoint or consequence.
Target–Disease Evidence	How strongly is the proposed target connected to the relevant disease process?	Human genetics, functional evidence, perturbation studies, disease biology, and convergent observations may support the target hypothesis.	Defines the biological rationale that an artificial intelligence model may help organize or prioritize.	Treating statistical association or model ranking as proof of causal therapeutic relevance.	Target–disease association does not alone establish mechanism, tractability, safety, or therapeutic benefit.
Mechanistic Plausibility	Is there a credible account of how modulation of the target or pathway could alter the disease-relevant process?	Mechanistic evidence may arise from molecular biology, pathway studies, perturbation experiments, and orthogonal validation.	Distinguishes predictive association from a biologically interpretable intervention hypothesis.	Inferring mechanism from feature importance, correlation, or post-hoc explanation.	Model explanation may support investigation but cannot independently confirm mechanism or causality.
Intervention Feasibility and Tractability	Can the biological hypothesis be tested and potentially modulated using an appropriate therapeutic or experimental modality?	Structural, chemical, biophysical, modality-specific, and assay evidence define feasible intervention routes.	Connects biological prioritization to the practical capacity to generate discriminating evidence.	Advancing a biologically attractive target despite the absence of a realistic intervention or measurement strategy.	Biological plausibility does not establish chemical, technical, or modality-specific tractability.
Chemical and Screening Evidence	Which compounds or chemical series should receive higher-fidelity computation or experimental testing?	Virtual screening, molecular-property prediction, docking, similarity analysis, and medicinal-chemistry knowledge may reduce the search space.	Defines a bounded use in which artificial intelligence can support prioritization rather than determine activity.	Equating a predicted score, rank, or generated structure with confirmed binding, activity, or developability.	Computational prioritization requires subsequent physical or experimental evaluation.
Safety and Liability Evidence	What plausible harms, off-target effects, assay artefacts, or developability liabilities could change the decision?	Safety biology, selectivity evidence, chemical alerts, assay controls, exposure considerations, and uncertainty analysis inform risk.	Makes the evidence burden proportional to the consequences of false reassurance or erroneous exclusion.	Using incomplete or prediction-only safety evidence to make irreversible progression decisions.	Absence of a predicted liability is not evidence that a liability is absent.
Evidence Integration and Provenance	How were heterogeneous observations, predictions, assumptions, and expert judgments combined?	Traceable provenance, source quality, evidence concordance, and explicit contradiction handling support auditability.	Prevents composite scores from concealing weak evidence, double counting, or incompatible assumptions.	Treating numerical aggregation as objective proof or allowing dependent sources to appear independent.	Integration can organize evidence but cannot create information that is absent from the underlying sources.
Experimental Resolvability	Can remaining uncertainty be reduced through feasible synthesis, assay, perturbation, or orthogonal confirmation?	Early discovery progresses through experiments designed to discriminate among competing hypotheses.	Links model outputs to an evidence-generation strategy rather than a terminal prediction.	Allowing a model recommendation to persist without specifying how it could be challenged or falsified.	A model is not decision-ready for consequential use when its critical claims cannot be tested or monitored.
Human and Organizational Accountability	Who reviews the output, challenges assumptions, authorizes action, and records the reasoning?	Multidisciplinary judgment is required to interpret biological, chemical, technical, and portfolio evidence.	Establishes responsibility for model influence and enables escalation, override, and re-evaluation.	Automation bias, rubber-stamping, or diffusion of responsibility across teams.	Human oversight does not compensate for invalid data, inadequate validation, or unsupported claims.

Why Model Performance Alone Cannot Establish Decision Readiness

Retrospective model performance is an incomplete basis for decision readiness because apparent gains may reflect memorization of closely related compounds, exploitable chemical-data biases, or evaluation and reporting practices that fail to reveal optimistic analytical choices [8-10]. These problems are especially consequential in pharmaceutical applications, where observations are rarely independent and identically distributed in the simple statistical sense. Compounds share scaffolds, assays vary in protocol and quality, labels may encode laboratory-specific practices, and datasets are often assembled from sources with different inclusion rules. A high score can therefore indicate genuine transferable learning, local interpolation within a familiar chemical series, recognition of dataset artefacts, or some combination of these. Without an evaluation design that distinguishes among those possibilities, the result cannot determine the degree of influence the model should have.

Data quantity does not resolve this problem when the available observations are poorly matched to the intended decision. Chemical and biological datasets can be large while remaining sparse in the most relevant region, inconsistent across assays, biased toward historically studied targets, or insufficiently representative of future compounds and experimental conditions [11]. The distinction between data availability and data suitability is therefore central to decision readiness. Suitability asks whether the observations support the endpoint, intervention, chemical space, biological context, and time horizon relevant to the proposed use. It also asks whether the label corresponds to the decision-relevant phenomenon or merely to a convenient proxy. A model trained on unsuitable data may be statistically stable and reproducible while still being scientifically uninformative for the decision.

The same limitation applies to commonly reported metrics. Discrimination, ranking, calibration, enrichment, prediction error, and generative-objective scores answer different questions and may change under different splitting strategies, comparators, or prevalence conditions. A model may rank compounds effectively but provide poorly calibrated probabilities; it may predict an assay endpoint accurately while failing on new scaffolds; or it may generate formally valid structures that are not synthesizable, selective, safe, or therapeutically relevant. Performance must therefore be interpreted as a structured profile rather than a single scalar property. The relevant profile includes the comparator, evaluation distribution, error types, sensitivity to analytical choices, stability across model versions, and the relationship between the measured endpoint and the downstream pharmaceutical action.

Decision readiness additionally requires evidence that a model improves or appropriately constrains an existing decision process. This is a stronger claim than showing that one model outperforms another on a benchmark. The model may need to identify compounds for confirmatory testing, expose uncertainty that changes experimental sequencing, prevent an avoidable error, or integrate evidence more consistently than the current workflow. Whether such value exists depends on the decision context, the cost and reversibility of errors, available alternatives, and the capacity of users to interpret and challenge the output. The proposed decision-readiness model therefore treats technical performance as one evidence component within a broader justification. Performance can support readiness, but it cannot establish readiness alone.

The Proposed Decision-Readiness Model: Context, Risk, Evidence, and Oversight

The proposed decision-readiness model defines readiness as a property of a specific model–decision relationship rather than as a stable attribute of an algorithm. Its first component, decision context, specifies the pharmaceutical choice that may be influenced, the intended users, the relevant biological or chemical domain, the endpoint, the available alternatives, and the time horizon over which the output is expected to remain informative. Context also determines the legitimate role of model explanations. Explainable artificial intelligence may help scientists inspect learned relationships, formulate hypotheses, identify unexpected dependencies, or challenge a prediction, but an apparently plausible explanation does not independently establish biological mechanism, causality, or experimental truth [12]. A model cannot therefore be declared ready merely because its outputs are interpretable. The explanation must be relevant to the decision, sufficiently faithful to the model, and used within a process capable of distinguishing computational rationale from scientific confirmation.

The second component, consequence and risk, concerns what may happen if the model is wrong, unstable, misunderstood, or used beyond its supported domain. Risk is not limited to direct safety harm. In early discovery, erroneous model influence may divert experiments, suppress a valid hypothesis, promote a misleading chemical series, consume scarce synthesis capacity, or distort portfolio allocation. Responsible machine-learning principles support increasing the burden of validation and governance as the consequences of model-guided action become more substantial [13]. The proposed model accordingly treats risk as a determinant of evidence requirements rather than as a final warning appended after technical development. Low-consequence, reversible exploratory use may tolerate greater uncertainty, whereas exclusionary, progression, or portfolio-influencing decisions require stronger evidence, explicit alternatives, and reliable opportunities for challenge and reversal.

The third component, evidence adequacy, asks whether the data, validation design, experimental corroboration, and provenance support the exact claim being made. The fourth, human and organizational oversight, identifies who is responsible for interpreting the output, challenging assumptions, authorizing action, recording dissent, and initiating revalidation. These components are interdependent. Strong retrospective performance cannot compensate for a poorly defined decision, and extensive human review cannot repair unsuitable data or invalid evaluation. Evidence of technical capability may also fail to create value when the model is misaligned with workflow, users, external conditions, or accountability arrangements [14]. Oversight must therefore be substantive rather than ceremonial. Reviewers need access to the model's intended-use statement, limitations, uncertainty, supporting evidence, relevant alternatives, and known failure modes. They must also possess the authority to defer, override, or restrict use.

The proposed model does not combine context, risk, evidence, and oversight into a universal numerical score. A single composite value could conceal a critical deficiency, create false precision, or allow strength in one component to offset an unacceptable weakness in another. Instead, readiness is treated as conditional on the adequacy of every component that is critical to the proposed use. Cross-cutting controls for leakage, reproducibility, versioning, uncertainty, transportability, and failure-mode detection constrain the entire assessment. Leakage and hidden analytical dependencies can produce overoptimistic scientific claims even when the reported workflow appears technically sophisticated, making auditability a prerequisite for consequential model influence [15]. **Figure 1** presents the pharmaceutical decision-readiness airlock, showing how the article's key components and boundaries are connected within the proposed decision-readiness model: context, risk, evidence, and oversight.

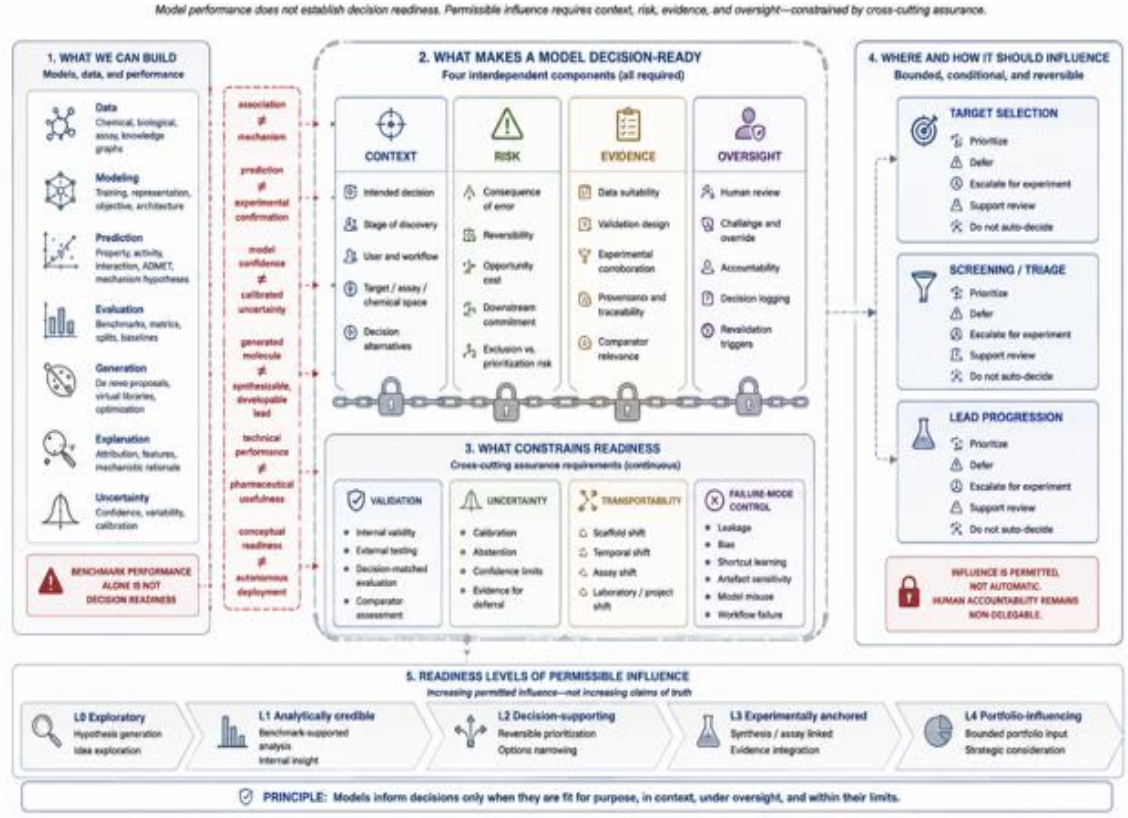

Figure 1. Pharmaceutical Decision-Readiness Airlock.

The figure is an original conceptual synthesis that organizes the article's central contribution across decision points in early pharmaceutical discovery and their evidentiary demands, model performance alone cannot establish decision readiness, proposed decision-readiness model: context, risk, evidence, and oversight, proposed decision-readiness model, readiness levels from exploratory analysis to portfolio-influencing use, validation, uncertainty, transportability, and failure-mode requirements. 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.

Readiness Levels from Exploratory Analysis to Portfolio-Influencing Use

The proposed readiness levels classify the degree of influence that a model may permissibly exert rather than ranking algorithms by sophistication. Level 0, exploratory use, permits a model to identify patterns, generate hypotheses, compare representations, or reveal potentially informative regions of chemical or biological space without directly changing a project decision. This level is especially relevant to generative systems because molecular and protein-generation tasks differ in objectives, representations, constraints, and evaluation practices; consequently, a model's apparent success in one generative setting cannot be converted into a universal readiness designation [16]. At Level 0, outputs should be described as computational proposals or analytical observations. They may guide discussion or method development, but they should not be presented as confirmed candidates, mechanisms, or decision-ready recommendations.

Level 1, analytically credible use, requires transparent data provenance, appropriate comparators, leakage-resistant evaluation, reproducible implementation, and evidence that the model measures what it claims to measure. Standardized benchmarks can support this level by enabling controlled comparisons of distribution learning, optimization, and goal-directed generation [17]. Nevertheless, benchmark success establishes only bounded analytical credibility. It does not show that the generated or prioritized outputs satisfy the full set of pharmaceutical constraints relevant to synthesis, assay behaviour, selectivity, developability, or therapeutic value. A Level 1 model may inform technical model selection, sensitivity analysis, or retrospective prioritization exercises, but consequential actions should not depend on its output without additional decision-specific evidence.

Level 2, decision-supporting use, permits a model to influence reversible prioritization, such as selecting compounds for higher-fidelity computation, proposing experiments, or ordering candidates for expert review. At this level, the output must be linked to a defined decision rule, an applicability domain, calibrated uncertainty or an appropriate limitation signal, and an explicit human-review process. For generative models, syntactic validity, novelty, or optimization of a computational objective is insufficient because proposed molecules may remain difficult or impossible to synthesize [18]. Level 2 therefore requires assessment of whether the model's recommendations can be acted upon and challenged through feasible follow-up. The model may narrow options or identify evidence gaps, but it must not collapse molecular plausibility, synthetic accessibility, biological activity, and developability into one undifferentiated claim.

Level 3, experimentally anchored use, requires that model-guided proposals be connected to synthesis, assay, perturbation, or another relevant physical test. Experimental coupling increases evidentiary maturity because it exposes computational proposals to constraints not fully represented in training data or proxy objectives. Work integrating generative artificial intelligence with on-chip synthesis illustrates how design, production, and testing can be joined within an iterative evidence-generating process [19]. Even so, a successful proof-of-concept remains bounded by its target, assay, chemistry, and experimental design. Level 4, portfolio-influencing use, is the highest proposed level and concerns decisions that materially alter program continuation, comparative investment, or resource allocation. It requires accumulated validation, experimentally supported project evidence, explicit uncertainty, documented challenge, and senior accountable oversight. Level 4 does not mean that the model controls the portfolio; it means that its contribution is sufficiently characterized to serve as one traceable input to an accountable decision.

Validation, Uncertainty, Transportability, and Failure-Mode Requirements

Decision-ready validation must evaluate more than average predictive performance. Molecular uncertainty-estimation methods differ in calibration, error discrimination, scalability, and suitability for guiding additional evidence generation; they must therefore be assessed for the intended endpoint, chemical domain, distribution shift, and action policy rather than accepted on the basis of method name alone [20-22]. A useful uncertainty estimate should help distinguish cases in which the model may support action from cases requiring abstention, expert escalation, or further experimentation. Model confidence is not automatically calibrated uncertainty, and calibrated uncertainty on one evaluation distribution may fail under another. The proposed model therefore requires the intended use of uncertainty to be specified in advance: communicating limits, ranking evidence needs, triggering deferral, or guiding acquisition.

Transportability concerns whether a model remains informative when the chemistry, assay, target, laboratory, time period, or project setting differs from the data used for development. A single random or held-out test split cannot represent all such changes. Evaluations across controlled levels of train–test relatedness demonstrate that apparent molecular generalizability can vary substantially depending on how closely evaluation examples resemble the training data [23]. Transportability should consequently be examined through shift-specific tests selected to approximate the intended use. Scaffold-aware, temporal, laboratory-specific, target-specific, or externally sourced evaluations may each answer different questions. Failure in one setting does not make a model universally useless, but it restricts the domain within which decision influence can be justified.

Failure-mode requirements convert general caution into explicit tests. Relevant failures include data leakage, duplicated or dependent observations, assay artefacts, label errors, confounding by chemical series, shortcut learning, unstable rankings, miscalibrated uncertainty, extrapolation beyond the training domain, implausible explanations, unrealistic generated structures, and workflow-induced automation bias. The assessment should identify which failures are plausible, how they would be detected, and what response they would trigger. High-impact modes require deliberate stress testing rather than passive monitoring. Failure-mode analysis must also include the surrounding workflow because a technically valid prediction may still be misused when uncertainty is hidden, outputs are presented without alternatives, or reviewers lack the authority to challenge recommendations.

The resulting validation package should be proportional to the intended readiness level. Exploratory use may rely on reproducibility, clear provenance, and bounded analytical evaluation. Decision-supporting use additionally requires decision-matched comparators, applicability analysis, uncertainty linked to an action rule, and documentation of human review. Experimentally anchored use requires prospective physical or biological evidence and analysis of discordance between computational and experimental findings. Portfolio-influencing use requires accumulated evidence across model versions and project stages, explicit consideration of opportunity costs, and controlled revalidation when data, assays, targets, or decision processes change. No finite package can eliminate uncertainty. The purpose of validation is to define and reduce uncertainty sufficiently for a bounded use, not to certify universal correctness.

Operational Use in Target Selection, Screening, and Lead Progression

In target selection, artificial intelligence may help organize heterogeneous evidence, detect patterns across sources, identify contradictions, or rank hypotheses for deeper review. Genetic support is one potentially valuable input, but estimates of its relationship with downstream approval depend on how targets, mechanisms, indications, and evidence are defined [24]. Operational readiness therefore requires a model to show how genetic information contributes to the target case rather than converting it into a deterministic target score. The decision record should separate observed genetic evidence from inferred mechanism, state what additional biological or safety evidence is needed, and reveal whether the target ranking is stable under alternative assumptions. At lower readiness levels, the model may guide evidence collection. At higher levels, it may influence comparative prioritization only when the broader target case is independently reviewable.

In virtual screening, the operational decision is usually not whether a compound is a drug, but which compounds should receive more expensive computation or physical testing. Deep-learning-assisted docking workflows demonstrate how surrogate models can reduce the computational burden of searching very large libraries by prioritizing compounds for subsequent higher-fidelity docking [25]. This use is decision-relevant because the model allocates screening resources, yet its inference remains bounded. A high surrogate or docking score does not establish binding, functional activity, selectivity, or developability. Readiness depends on the library, target representation, chemical domain, validation split, enrichment objective, uncertainty,

and planned confirmatory process. A screening model may be ready for triage while remaining wholly unready to support a lead-progression claim.

Experimental testing provides an essential boundary between virtual prioritization and confirmed activity. Ultra-large library docking followed by physical testing has shown how computational ranking can nominate previously unexplored chemotypes for experimental evaluation [26]. The model's operational value lies in narrowing the search space and structuring the order of testing, not in eliminating the assay. Decision readiness in this context therefore requires a documented hand-off: which compounds are selected, why they are selected, what uncertainty or diversity criteria are applied, which assays will test the relevant claim, and how negative or discordant results will update the model or decision. Without this hand-off, a virtual-screening result remains an analytical output rather than an experimentally grounded pharmaceutical decision aid.

Lead progression creates a higher evidentiary burden because advancement usually reflects multiple properties rather than a single predicted endpoint. Artificial intelligence-assisted design followed by synthesis and biochemical testing can support lead-generation choices by demonstrating that computational proposals are physically realizable and capable of producing relevant experimental activity [27]. Such evidence does not establish selectivity, exposure, safety, manufacturability, clinical utility, or ultimate therapeutic value. A model may therefore support progression from proposal to tested hit while remaining insufficient for later advancement. Operational readiness requires explicit separation of predicted properties from measured properties, documentation of unresolved liabilities, assessment of series-level evidence, and multidisciplinary review. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop operational use in target selection, screening, and lead progression within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from When Is an Artificial Intelligence Model Ready to Influence Decisions in Early-Stage

Component or Layer	Inputs	Core Function	Expected Output	Uncertainty or Limitation	Validation Requirement
Decision-Context Specification	Intended choice, users, endpoint, stage, alternatives, time horizon, reversibility	Define the exact model–decision relationship	Bounded intended-use statement and explicit prohibited uses	Context may change as the project, assay, or evidence base evolves	Independent review that the evaluated task matches the intended pharmaceutical decision
Consequence And Risk Assessment	Scientific consequences, resource commitments, opportunity costs, downstream exposure, reversibility	Determine the required burden of evidence and oversight	Risk-proportionate validation and escalation plan	Organizational consequences may be difficult to quantify and may differ among programs	Scenario analysis and documented agreement on unacceptable errors and override conditions
Evidence-Adequacy Layer	Data provenance, biological evidence, chemical evidence, validation results, experimental findings	Determine whether the specific decision claim is supported	Traceable evidence profile with gaps and contradictions	Evidence may be indirect, incomplete, dependent, or affected by hidden bias	Source-quality assessment, leakage controls, comparator analysis, and stage-appropriate corroboration
Uncertainty and Abstention Layer	Predictive distributions, calibration evidence, applicability signals, model disagreement	Identify when output should support action, trigger more evidence, or be deferred	Decision-linked uncertainty statement and abstention rule	Uncertainty estimates may be miscalibrated or incomplete under distribution shift	Calibration and error-retention evaluation under conditions approximating intended use
Transportability Layer	Chemical-space coverage, assay context, target class, laboratory and temporal information	Bound the environments in which model influence is justified	Explicit applicability domain and shift-specific restrictions	Future projects may differ in ways not represented by available external tests	Scaffold, temporal, laboratory, target, or external validation selected for the proposed use
Failure-Mode Control	Leakage taxonomy, label-quality analysis, artefact tests, perturbations, edge cases, workflow review	Detect plausible technical and sociotechnical mechanisms of misleading output	Failure register, detection procedures, and mitigation or escalation actions	Unknown failure modes cannot be eliminated through a finite test set	Prespecified stress testing, independent audit, version comparison, and post-use surveillance
Human and Organizational Oversight	Named reviewers, decision rights, multidisciplinary expertise, decision logs	Make model influence challengeable, reversible, and accountable	Recorded review, dissent, override, and authorization	Human review may be affected by automation bias or institutional pressure	Evaluation of whether reviewers can detect errors, understand limits, and exercise meaningful control
Target-Selection Use	Genetics, disease biology, perturbation evidence, safety, tractability, strategic fit	Organize evidence and prioritize hypotheses for additional validation	Comparative target case and explicit evidence-generation plan	Association may be mistaken for mechanism or therapeutic causality	Independent biological review and prospective testing of decision-relevant target hypotheses
Screening and Hit-Triage Use	Target representation, chemical library, assay information, model scores, diversity and uncertainty criteria	Prioritize compounds for higher-fidelity computation or physical testing	Ranked and bounded test set with selection rationale	Scores may capture artefacts, similarity, or proxy objectives rather than activity	Confirmatory assays, orthogonal testing, artefact controls, and analysis of false exclusions

Lead-Progression Use	Measured potency, selectivity, ADME, safety, synthesis, series evidence, predicted liabilities	Integrate multi-parameter evidence and expose trade-offs	Documented progression recommendation with unresolved uncertainties	Predicted developability may be mistaken for demonstrated developability	Cross-functional review and stage-appropriate experimental confirmation before advancement
Portfolio-Influencing Layer	Project evidence, readiness record, strategic constraints, opportunity costs, model history	Support comparative resource allocation without automating accountability	Traceable contribution to an accountable portfolio decision	Long outcome lags and confounding make model-specific value difficult to isolate	Prospective organizational evaluation, periodic revalidation, and audit of model influence on decisions
Research and Learning Layer	Model versions, decision records, experimental outcomes, disagreements, overrides, failures	Convert use into controlled evidence for model and process improvement	Versioned evidence base for revalidation and readiness reassessment	Outcome attribution may remain uncertain in multifactorial discovery programs	Longitudinal analysis of decisions, errors, corrections, and incremental value relative to existing practice

Limitations and Boundary Conditions

The proposed decision-readiness model is a conceptual synthesis and has not been empirically validated as a maturity scale, predictive instrument, or universally reliable governance standard. Its components are intended to organize evaluation rather than to generate a definitive readiness score. Terms such as context adequacy, acceptable risk, sufficient evidence, and meaningful oversight require operational definitions that may differ across organizations, therapeutic areas, modalities, data environments, and discovery stages. The model may improve the transparency of readiness claims, but it cannot establish that two independent reviewers or organizations will assign the same readiness level without further methodological development and inter-rater evaluation.

The supporting evidence base is also heterogeneous. It includes critical perspectives, methodological evaluations, validation recommendations, databases, benchmark studies, modelling analyses, and experimentally anchored examples. These sources illuminate different parts of the readiness problem, but they do not constitute direct validation of the proposed relationships. Evidence drawn from clinical artificial intelligence or general scientific machine learning has been used only to support broad principles concerning risk, workflow, leakage, accountability, and external validity; those principles require adaptation before they can be operationalized in early pharmaceutical discovery. Similarly, examples of successful screening, synthesis, or experimental testing demonstrate possible evidence pathways but do not establish that the same workflows, models, or thresholds will transport to other targets or programs.

The model must not be used to imply regulatory acceptance, clinical validity, guaranteed discovery success, or permission for autonomous decision-making. A high proposed readiness level does not mean that a candidate is safe, effective, developable, or suitable for clinical testing. It indicates only that the evidence supporting a model's influence on a specified early-discovery decision is stronger and more structured than at lower levels. Readiness is conditional, time-limited, and reversible. Changes in data provenance, model architecture, endpoint definition, target class, chemical space, assay protocol, organizational workflow, or consequence profile may invalidate a previous assessment and require revalidation.

Research Agenda and Implementation Priorities

The first research priority is to test whether decision-centred evaluation produces more reliable or useful conclusions than conventional model-centred benchmarking. This requires prospective studies in which the intended decision, comparator workflow, permissible degree of model influence, error consequences, and follow-up evidence are specified before model evaluation. Studies should record model versions, predictions, uncertainty, human judgments, selected actions, experimental outcomes, overrides, and disagreements. The central question is not simply whether the model predicts accurately, but whether its inclusion changes the quality, efficiency, consistency, or scientific defensibility of a decision relative to existing practice. Such studies should also examine whether readiness-level assignments are reproducible across independent reviewers and whether assigned levels predict later patterns of error, correction, or utility.

A second priority is the development of shared validation infrastructure for distribution shift, uncertainty, and failure modes in pharmaceutical settings. This includes reference protocols for scaffold, temporal, target, assay, laboratory, and data-source shifts; standard reporting of label generation and provenance; leakage audits; error-retention and abstention analyses; and structured case libraries documenting failed or misleading models. Infrastructure should support versioned evidence rather than static benchmark snapshots. It should also permit negative findings, failed transportability, and model–experiment discordance to remain visible. Without this infrastructure, readiness assessments may repeatedly rediscover the same limitations while favourable retrospective results remain easier to publish and reuse than evidence of failure.

The third priority is implementation research on human–AI decision systems. Pharmaceutical scientists, medicinal chemists, computational researchers, assay specialists, toxicologists, and portfolio leaders may interpret the same output differently because they hold different responsibilities and evidence standards. Studies should test which explanations help users detect errors, how uncertainty affects prioritization, when automation bias emerges, and which escalation structures enable meaningful challenge. Early implementation should begin with reversible, well-instrumented decisions and explicit prohibited uses. Progression to experimentally anchored or portfolio-influencing applications should occur only after the organization

can demonstrate traceable review, reliable hand-offs to evidence generation, controlled revalidation, and the ability to suspend model use when assumptions no longer hold.

Conclusion

An artificial intelligence model is ready to influence an early pharmaceutical-discovery decision only when its technical performance is embedded within a decision-specific justification. The proposed decision-readiness model organizes that justification through four interdependent components: context, consequence and risk, evidence adequacy, and human and organizational oversight. It further restricts model influence through readiness levels ranging from exploratory analysis to portfolio-influencing use, with validation, calibrated uncertainty, transportability, and failure-mode control operating across all levels. The central contribution is not a universal score or certification mechanism, but a disciplined distinction between what a model can compute and what available evidence warrants decision-makers doing with its output. The model remains conceptual, context-dependent, and subject to prospective testing. It cannot establish biological mechanism, experimental confirmation, developability, therapeutic value, regulatory acceptability, or deployment readiness. Its value will depend on whether future studies can operationalize its components, test the reproducibility of readiness assignments, and determine whether bounded model influence improves scientific reasoning and resource allocation without concealing uncertainty or displacing accountable judgment.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. 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
2. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
3. Schneider G. Automating drug discovery. Nat Rev Drug Discov. 2018;17(2):97-113. doi:10.1038/nrd.2017.232
4. Emmerich CH, Martinez Gamboa L, Hofmann MCJ, Bonin-Andresen M, Arbach O, Schendel P, et al. Improving target assessment in biomedical research: The GOT-IT recommendations. Nat Rev Drug Discov. 2021;20(1):64-81. doi:10.1038/s41573-020-0087-3
5. Plenge RM. Priority index for human genetics and drug discovery. Nat Genet. 2019;51(7):1073-5. doi:10.1038/s41588-019-0460-5
6. Hingorani AD, Kuan V, Finan C, Kruger FA, Gaulton A, Chopade S, et al. Improving the odds of drug development success through human genomics: Modelling study. Sci Rep. 2019;9(1):18911. doi:10.1038/s41598-019-54849-w
7. Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: Supporting systematic drug-target identification and prioritisation. Nucleic Acids Res. 2021;49(D1). doi:10.1093/nar/gkaa1027
8. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403
9. Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712
10. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, Harrow J, et al. DOME: Recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-7. doi:10.1038/s41592-021-01205-4
11. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
12. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84. doi:10.1038/s42256-020-00236-4
13. 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
14. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
15. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804

16. Tang X, Dai H, Knight E, Wu F, Li Y, Li T, et al. A survey of generative AI for de novo drug design: New frontiers in molecule and protein generation. Brief Bioinform. 2024;25(4). doi:10.1093/bib/bbae338
17. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096-108. doi:10.1021/acs.jcim.8b00839
18. Gao W, Coley CW. The synthesizability of molecules proposed by generative models. J Chem Inf Model. 2020;60(12):5714-23. doi:10.1021/acs.jcim.0c00174
19. Grisoni F, Huisman BJH, Button AL, Moret M, Atz K, Merk D, et al. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. Sci Adv. 2021;7(24). doi:10.1126/sciadv.abg3338
20. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502
21. Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975
22. Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci. 2021;7(8):1356-67. doi:10.1021/acscentsci.1c00546
23. Ektefaie Y, Shen A, Bykova D, Marin MG, Zitnik M, Farhat M. Evaluating generalizability of artificial intelligence models for molecular datasets. Nat Mach Intell. 2024;6(12):1512-24. doi:10.1038/s42256-024-00931-6
24. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet. 2019;15(12). doi:10.1371/journal.pgen.1008489
25. Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, Norinder U, et al. Deep Docking: A deep learning platform for augmentation of structure-based drug discovery. ACS Cent Sci. 2020;6(6):939-49. doi:10.1021/acscentsci.0c00229
26. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, et al. Ultra-large library docking for discovering new chemotypes. Nature. 2019;566(7743):224-9. doi:10.1038/s41586-019-0917-9
27. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40. doi:10.1038/s41587-019-0224-x