Pharmacophore

ISSN-2229-5402

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

WHERE ARTIFICIAL INTELLIGENCE CHANGES DRUG DISCOVERY—AND WHERE IT DOES NOT: A CRITICAL REVIEW OF EVIDENCE, LIMITATIONS, AND TRANSLATIONAL VALUE

Sophie Williams[1]*, Javier Rodriguez[2], Elena De Vries[3]

1. ***Department of AI in Drug Discovery and Critical Appraisal, Faculty of Pharmaceutical Sciences, University of Amsterdam, Amsterdam, Netherlands.***
2. ***Department of Translational Value and Evidence Assessment, Faculty of Pharmacy, University of Aruba, Oranjestad, Aruba.***
3. ***Department of AI Limitations and Scientific Reasoning, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands.***

ARTICLE INFO

Received:
08 May 2024
Received in revised form:
16 September 2024
Accepted:
19 September 2024
Available online:
28 October 2024

Keywords: Artificial intelligence, Drug discovery, Molecular design, Target identification, Virtual screening, Translational validation

ABSTRACT

Artificial intelligence has become embedded in target identification, molecular representation, compound generation, virtual screening, preclinical prediction, and selected development workflows. Yet the significance of these applications remains difficult to judge because computational performance, experimental usefulness, translational value, and routine pharmaceutical readiness are frequently discussed as though they were equivalent outcomes. This critical review examines where artificial intelligence materially changes drug-discovery practice and where its contribution remains conditional, indirect, or unconfirmed. The review applies a stage-specific analytical approach that distinguishes data availability from data suitability, benchmark performance from prospective usefulness, molecular plausibility from developability, prediction from explanation or causality, and experimental confirmation from clinical translation. The evidence indicates that artificial intelligence has its most defensible value in bounded tasks involving large or structurally informative datasets, clearly specified objectives, rapid candidate prioritization, and experimental feedback. Examples include relational target analysis, focused molecular generation, structure-enabled screening, and the prioritization of compounds for assay testing. However, many reported gains remain partly attributable to dataset composition, automation, benchmark construction, chemical similarity, or evaluation choices rather than to a transferable algorithmic advantage. Evidence becomes progressively thinner when claims move from discovery-task acceleration to mechanistic validity, preclinical predictiveness, clinical benefit, regulatory acceptability, or improved portfolio success. The principal conceptual contribution of this review is a proposed value-and-limit interpretation in which claims are judged according to their application stage, validation setting, experimental confirmation, and permissible influence on pharmaceutical decisions. The synthesis is necessarily constrained by heterogeneous study designs, uneven reporting, selective publication of successful campaigns, and limited independent prospective evidence. Responsible adoption therefore requires evidence proportional to the consequence of the decision being supported rather than reliance on a generic designation of artificial-intelligence capability.

To Cite This Article: Williams S, Rodriguez J, De Vries E. Where Artificial Intelligence Changes Drug Discovery—and Where It Does Not: A Critical Review of Evidence, Limitations, and Translational Value. Pharmacophore. 2024;15(5):27-36. https://doi.org/10.51847/2zdDmjoiLu

Introduction

Artificial intelligence now influences how pharmaceutical scientists organize biological relationships, represent molecules, search chemical space, prioritize experiments, and interpret complex preclinical or development data. Across target discovery, molecular design, screening, and development, AI can change how hypotheses are generated and prioritized, but its pharmaceutical value remains conditional on problem definition, data quality, and integration with experimental and medicinal-chemistry workflows [1–3]. This distinction is fundamental. A model may reduce the time required to score compounds, detect patterns that are difficult to identify manually, or enlarge the number of hypotheses considered without improving the biological validity of a target, the developability of a molecule, or the probability that a program will produce a

Corresponding Author: Sophie Williams; Department of AI in Drug Discovery and Critical Appraisal, Faculty of Pharmaceutical Sciences, University of Amsterdam, Amsterdam, Netherlands. E-mail: sophie.williams@uva.nl.

useful medicine. The relevant question is therefore not whether AI can perform pharmaceutical tasks, but which tasks it changes in a scientifically and decision-relevant manner.

The interpretation of AI performance is complicated by the nature of pharmaceutical data. Chemical and biological datasets may be large yet remain unsuitable for the intended prediction because their labels reflect heterogeneous assays, historical project choices, measurement noise, incomplete negative results, restricted chemical series, or unrecorded experimental context [4]. Models trained on such data can learn stable statistical regularities without learning relationships that remain valid for new targets, new chemical matter, different laboratories, or downstream development conditions. Data abundance must consequently be separated from data relevance, representativeness, and decision suitability. The availability of millions of structures or observations does not by itself establish that the resulting model can support causal inference, prospective experimentation, or a high-consequence pharmaceutical decision.

A further difficulty is the breadth of activities placed under the label of artificial intelligence. Applications range from literature mining and network analysis to molecular-property prediction, de novo design, image analysis, toxicity modeling, drug repurposing, trial support, and regulatory-facing analyses [5]. These applications do not share one evidentiary standard. A retrospective classifier, a model used to rank compounds for laboratory testing, a system that contributes to a synthesized active molecule, and a method used within a clinical-development submission occupy fundamentally different evidence positions. Treating them as comparable examples of "AI success" conceals differences in data provenance, validation design, experimental confirmation, uncertainty, and the consequence of error.

This review addresses that conceptual and evidentiary gap by evaluating artificial intelligence as a set of stage-specific pharmaceutical capabilities rather than as a single technological intervention. Task-level prediction and demonstrable improvement in drug-discovery productivity are not interchangeable claims [3]. The review therefore distinguishes demonstrated capability, benchmark performance, plausible utility, experimental confirmation, translational value, and routine readiness. Its original contribution is a proposed value-and-limit interpretation that maps what AI can credibly contribute at each stage, identifies recurrent pathways to overclaiming, and specifies the additional evidence required before a model should influence increasingly consequential scientific or development decisions. The proposed interpretation is a critical synthesis of the selected literature; it is not presented as an empirically validated maturity scale, regulatory standard, or universal decision framework.

Review Scope, Analytical Questions, and Critical Evaluation Approach

The review was designed as a critical and interpretive synthesis rather than a claim of systematic exhaustiveness. Its scope includes peer-reviewed evidence concerning AI-enabled target discovery, molecular design, virtual and high-throughput screening, preclinical prediction, computational pharmacology, clinical-development support, and regulatory-facing use. Methodological and evaluative sources were also included when they directly informed the interpretation of model validity, benchmark design, reproducibility, or translational evidence. In accordance with methodological guidance for literature reviews, the analytical purpose determined the source-selection logic, evidence extraction, comparison, and synthesis strategy [6]. Publications were therefore not treated as equivalent units to be summarized sequentially. They were examined according to the pharmaceutical claim they supported, the type of evidence provided, the setting in which validation occurred, and the strength of the inference that could reasonably follow.

The review was organized around seven connected questions. First, what constitutes an appropriate scope and critical-evaluation logic for judging AI in pharmaceutical research? Second, where do target-discovery, molecular-design, and screening methods produce capabilities that extend beyond conventional manual or computational practice? Third, how far do these capabilities extend into preclinical pharmacology, safety, clinical development, and regulatory use? Fourth, when are apparent gains better explained by data scale, automation, chemical similarity, or benchmark construction? Fifth, what limits remain in causality, generalization, uncertainty, external validation, and experimental confirmation? Sixth, how should evidence maturity and overclaiming risk be mapped across stages? Seventh, what forms of validation and governance are proportionate to the decisions that AI may influence? Although the review was not framed as a formal systematic review, transparent reporting principles were used to make the source-identification logic, screening boundaries, and synthesis claims interpretable [7].

Critical appraisal focused on six dimensions: directness of evidence, data and task suitability, evaluation realism, reproducibility, experimental confirmation, and translational relevance. Directness concerned whether a source supported the precise pharmaceutical claim for which it was used. Data and task suitability addressed whether model inputs, labels, and prediction targets corresponded to the intended use. Evaluation realism considered whether testing occurred on familiar retrospective data, externally held data, prospective candidates, or experimentally generated outcomes. Reproducibility included the transparency of data, code, model specification, and analytical choices. Experimental confirmation distinguished computational predictions from synthesized, assayed, or otherwise independently tested outputs. Translational relevance asked whether the evidence supported a scientific hypothesis, an experimental decision, a development decision, or a stronger clinical or regulatory inference. These dimensions adapt established appraisal principles concerning methodological transparency, bias, and warranted conclusions without implying that an intervention-focused appraisal instrument has been directly validated for pharmaceutical AI reviews [8]. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop review scope, analytical questions, and critical evaluation approach within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Review scope, analytical questions, and critical evaluation approach

Evidence Domain	Eligible Evidence or Method	Reported Capability or Claim	Strength of Support	Reproducibility or Bias Concern	Unresolved Gap
Review Scope And Evidence Selection	Peer-reviewed experimental studies, benchmark analyses, methodological evaluations, critical reviews, validation guidance, and regulatory-facing landscape analyses	The evidence base can be organized by pharmaceutical stage and claim type rather than by algorithm family alone	Strong for structuring a critical review; not evidence of effectiveness by itself	Selective publication, heterogeneous terminology, and uneven reporting can shape the apparent evidence landscape	A shared reporting vocabulary for AI contribution, validation setting, and decision influence remains underdeveloped
Target and Biological Hypothesis Generation	Graph learning, knowledge integration, omics-based prediction, literature mining, and relational biomedical modeling	AI can prioritize targets or biological relationships for further investigation	Conditional; strongest for ranking and pattern discovery	Curated knowledge may reproduce literature bias, database incompleteness, and historical research concentration	Independent perturbational and mechanistic confirmation is often absent
Molecular Generation and Optimization	Latent-variable models, chemical language models, graph generators, reinforcement learning, and property-directed optimization	AI can generate or prioritize molecules satisfying modeled objectives	Strong for computational generation; more limited for pharmaceutical value	Training-set similarity, objective misspecification, invalid structures, and selective reporting of favorable candidates	Synthesizability, multiparameter developability, safety, and therapeutic relevance require separate testing
Screening and Property Prediction	Ligand-based models, structure-based models, virtual screening, image analysis, and high-throughput experimental learning	AI can rank compounds or predict defined molecular and biological endpoints	Variable; can be strong for bounded tasks with appropriate external or prospective testing	Leakage, scaffold similarity, assay artifacts, class imbalance, and inappropriate data splitting can inflate performance	Transportability to new chemistry, targets, laboratories, and decision contexts remains inconsistent
Experimental Confirmation	Prospective compound selection, synthesis, biochemical assays, cellular assays, animal studies, or independent laboratory replication	Model-prioritized outputs retain value when tested outside the training and benchmark environment	Stronger than retrospective evidence but still campaign-specific	Positive-result selection, incomplete reporting of failed candidates, and lack of independent replication	Few studies connect initial confirmation to sustained lead optimization or development progression
Translational And Decision Value	Preclinical safety assessment, clinical-development support, regulatory submissions, and longitudinal portfolio evidence	AI may influence development decisions beyond early discovery	Emerging and highly context-dependent	Confounding by expert judgment, automation, proprietary data, organizational capability, and changing workflows	Direct evidence that AI improves patient outcomes, regulatory conclusions, or portfolio success remains limited

Table note: Strength-of-support descriptions are qualitative review interpretations. They are not numerical scores, universal readiness classifications, or regulatory thresholds.

AI Contributions to Target Discovery, Molecular Design, and Screening

AI changes early discovery most clearly when it expands the scale or structure of hypothesis generation while preserving a defined path to experimental testing. Graph-based learning can integrate molecular structures, proteins, interactions, phenotypes, diseases, and biomedical knowledge into relational representations that support target prioritization, interaction prediction, and compound–target inference [9]. Across graph-based target inference, generative molecular design, and deep-learning screening, AI can generate or prioritize testable discovery hypotheses, with the strongest demonstrations including experimental follow-up [9–11]. The important contribution is not that these methods discover biological truth autonomously, but that they can reorganize complex evidence and identify candidates that may not be prioritized through simpler similarity searches or manual inspection. Their outputs remain conditional on the completeness and validity of the underlying relationships. A graph model can rank a target because it occupies a statistically informative network position, but this does not establish that perturbing the target will produce a therapeutically useful effect.

Molecular design illustrates both the genuine change introduced by AI and the boundary of that change. Continuous molecular representations allow candidate structures to be encoded in spaces where modeled properties can be predicted and optimization procedures can search for combinations favored by the objective function [12]. This capability changes the mechanics of design by enabling rapid generation, interpolation, and prioritization across chemical spaces that cannot be enumerated manually. However, property-directed generation remains only as meaningful as the property model and the encoded constraints. A candidate can be novel, mathematically valid, or predicted to bind a target while remaining synthetically inaccessible, chemically unstable, promiscuous, toxic, poorly exposed, or unsuitable for formulation. The evidentiary position becomes stronger when generated candidates are synthesized and experimentally evaluated, as demonstrated in a focused kinase-inhibitor campaign [10]. Even then, the result establishes prospective task utility for a particular target and workflow rather than a general improvement in medicinal-chemistry productivity or clinical success.

Screening applications provide some of the clearest examples of AI contributing to experimentally consequential prioritization. Computational advances in protein-structure modeling, docking, molecular representation, and large-scale screening can enlarge the set of compounds and binding hypotheses available for investigation [13]. Their value lies principally in ranking or filtering: the model directs limited experimental resources toward candidates judged more informative or promising. In

antibiotic discovery, deep learning was used to prioritize compounds that were subsequently evaluated biologically, providing stronger evidence than a purely retrospective benchmark comparison [11]. Nevertheless, such evidence remains bounded by the endpoint, chemical libraries, biological systems, and experimental procedures of the campaign. A successful identification does not show that the same model will perform similarly for another pathogen, target class, assay technology, or chemical distribution.

The critical synthesis is therefore that AI changes target discovery, molecular design, and screening by altering search capacity, representational flexibility, and prioritization efficiency. Graph learning can connect evidence dispersed across relational biomedical data [9]. Generative modeling can formulate property-directed molecular hypotheses [12]. Structure-enabled computational methods can broaden the range of targets and compounds considered [13]. These contributions are scientifically important, but they occupy different evidence levels. Retrospective target ranking supports hypothesis generation; computational molecular generation supports candidate formulation; prospective screening followed by laboratory testing supports bounded experimental utility. None of these levels alone establishes causal target validity, multiparameter developability, safety, clinical effectiveness, or regulatory acceptability. The defensible claim is that AI can improve how early-discovery hypotheses are generated and selected for testing, provided that model outputs remain subordinate to chemical, biological, and experimental validation.

Applications in Preclinical Development, Pharmacology, and Clinical Translation

Beyond early discovery, artificial intelligence is being applied to toxicity prediction, phenotypic analysis, and clinical-trial design, but each application operates under a different evidentiary burden [14–16]. In preclinical safety, models may identify associations between chemical or biological descriptors and defined toxicity endpoints, helping to prioritize compounds, assays, or potential adverse effects. In high-content imaging, deep-learning methods can extract phenotypic representations from cellular images at a scale that would be difficult to achieve through manual feature engineering. In clinical-trial design, AI may assist with protocol planning, patient identification, recruitment, monitoring, or operational forecasting. These capabilities can improve information processing and prioritization, but they do not establish that the underlying biological endpoint is valid, that the prediction will transport across laboratories or populations, or that using the model will improve a development outcome.

Regulatory-facing use provides evidence that AI is moving beyond exploratory research into practical development settings. A landscape analysis of regulatory submissions showed that AI and machine-learning components have been used across a range of drug and biologic development activities [17]. This finding is important because it demonstrates actual engagement with regulated development processes rather than hypothetical future use. However, presence within a submission is not equivalent to acceptance of a general algorithm class, endorsement of a specific modeling strategy, or recognition that AI-generated evidence can replace conventional evidence. Regulatory assessment is context-specific and depends on the model's role, documentation, validation, uncertainty, and influence on the decision under review. The appropriate interpretation is therefore that AI has entered regulatory-facing workflows, not that it has achieved unrestricted regulatory readiness.

Clinical-outcome prediction represents a still more demanding use case because the endpoint reflects many interacting determinants, including target biology, molecule quality, patient selection, trial design, adherence, standard of care, and operational execution. Multimodal models have been developed to estimate clinical-trial outcomes from target-related and trial-design information, with reported validation across more than one evaluation setting [18]. Such models may support program-level risk stratification or identify factors associated with trial progression. Nevertheless, predicted trial outcome should not be conflated with prediction of individual patient benefit, causal confirmation of the target, or proof that a compound is effective and safe. Clinical-trial databases also encode historical sponsor behavior, indication selection, protocol conventions, and reporting practices. A model may therefore learn regularities of development history as well as properties of the therapeutic hypothesis.

The evidence across preclinical and clinical applications supports a graduated interpretation. AI can plausibly improve the efficiency with which toxicity signals, cellular phenotypes, trial characteristics, and development records are analyzed. Safety models can help prioritize risks but do not constitute universal safety determinations [14]. Imaging models can structure complex phenotypic observations but do not establish mechanism from image similarity alone [15]. Trial-design systems may support operational and analytical choices without demonstrating improved clinical outcomes [16]. Regulatory submissions show real-world use but not general acceptability [17]. Trial-outcome models may contribute to portfolio prioritization while remaining vulnerable to confounding and transportability limitations [18]. The downstream value of AI is consequently more conditional than its early-discovery computational value because the relevant outcomes become biologically, organizationally, and clinically more complex.

Where Reported Gains Reflect Data Scale, Automation, or Benchmark Design

Benchmarking is necessary for comparing molecular machine-learning methods, but it also defines what counts as success. MoleculeNet helped standardize datasets, task definitions, data splits, metrics, and baseline comparisons across molecular prediction problems [19]. Such standardization supports cumulative method development and makes computational results easier to compare. Yet every benchmark represents a bounded evaluation environment. Its labels, chemical coverage, class balance, split strategy, and metric determine which model behaviors are rewarded. A method that performs well within a benchmark can be considered effective under those specified conditions, but the result does not independently establish

performance on new project chemistry, under different assay procedures, or in a pharmaceutical decision with asymmetric costs of error.

Some reported gains can arise because a benchmark permits models to exploit similarity structures that would be weaker in genuine prospective use. Ligand-based classification benchmarks have been shown to reward memorization of dataset-associated patterns rather than transferable generalization in some settings [20]. Chemical and structural biases can likewise inflate apparent virtual-screening performance when active and inactive compounds differ in ways unrelated to the biological interaction that the model is expected to learn [21]. These findings do not imply that benchmark evaluation is inherently invalid. They show that benchmark superiority must be interpreted in relation to the data-generating process and the intended application. Random splits are particularly vulnerable when close analogues occur in both training and test sets, while scaffold, temporal, target-family, or externally held evaluations may provide more demanding tests of transfer.

Data scale and automation can also produce real operational gains that are easily attributed entirely to the learning algorithm. Computer-assisted synthesis planning illustrates how model behavior depends on the reaction datasets used for training and evaluation [22]. Larger datasets may improve coverage, but they may also reproduce the reaction preferences, reporting conventions, and omissions of the source literature or proprietary archive. Similarly, an AI-enabled workflow may appear superior because it connects database search, scoring, compound registration, experiment scheduling, or robotic execution more efficiently than a conventional fragmented workflow. That improvement is meaningful, but it should be described accurately as a joint property of the algorithm, data infrastructure, automation, and organizational process rather than evidence that the predictive model alone is scientifically superior.

Evaluation must therefore be aligned with intended use. Guidance for machine learning in the chemical sciences emphasizes appropriate baselines, realistic splits, applicability-domain analysis, uncertainty assessment, reproducible reporting, and comparison against methods that reflect existing practice [23]. A useful evaluation should determine not only whether a new model achieves a favorable metric but also whether it changes a decision, identifies useful candidates beyond simpler strategies, remains reliable under relevant distribution shift, and fails in recognizable ways. Reported gain can reflect genuine representational or algorithmic improvement, improved data curation, larger training resources, more extensive hyperparameter search, workflow automation, or a permissive benchmark. These sources of gain should be disentangled wherever possible because they imply different expectations for replication and translation.

Persistent Limits in Causality, Generalization, Validation, and Experimental Confirmation

Prediction, explanation, mechanism, and causality are distinct scientific claims. Explainable-AI methods may identify molecular fragments, features, examples, or counterfactual changes associated with a model's prediction, thereby helping researchers inspect model behavior or formulate hypotheses [24]. Such explanations can be scientifically useful when they expose implausible dependencies, support medicinal-chemistry interpretation, or suggest experiments. They do not, however, demonstrate that the highlighted feature causes the biological effect or that the model has recovered the operative mechanism. Post hoc explanations may be unstable, method-dependent, or faithful only to an approximate representation of the model. Mechanistic claims require independent biological evidence, and causal claims require designs capable of distinguishing intervention effects from association.

Uncertainty creates a related interpretive challenge. Deep molecular models can produce high-confidence predictions even for compounds that differ substantially from the training data. Comparative evaluation of scalable uncertainty-estimation methods has shown that their behavior varies across tasks and models [25]. An uncertainty score must therefore be assessed for calibration, ranking ability, sensitivity to distribution shift, and usefulness within a specified decision rule. Model confidence is not equivalent to calibrated pharmaceutical uncertainty. Even a well-calibrated predictive distribution may omit uncertainty arising from assay error, biological heterogeneity, target validity, experimental execution, or downstream development. Responsible use requires identifying which uncertainty sources are represented and which remain outside the model.

The central translational issue is predictive validity: whether a model, assay, or experimental system predicts the outcome that matters for the decision in which it is used. Predictive validity in drug discovery is context-dependent and cannot be inferred from technical sophistication or internal fit alone [26]. A model used to rank compounds for a biochemical assay requires evidence that its ranking improves that assay-level decision. A preclinical model used to support progression toward human testing requires evidence connecting its outputs to relevant human outcomes or to an accepted intermediate decision. Validation should therefore be indexed to context of use. The same algorithm can be adequate for low-consequence prioritization and inadequate for exclusion of a promising program, safety assessment, dose selection, or interpretation of clinical evidence.

Prospective experimental confirmation provides stronger evidence than evaluation on data available before model development. Prospectively validated proteochemometric models for small-molecule binding to bromodomain proteins illustrate how predictions can be tested against new compounds and protein contexts rather than only against a retrospective split [27]. Such studies evaluate whether the model supports a real selection before the relevant result is known. Prospective confirmation nevertheless remains bounded. It may establish performance for a target family, assay, chemical domain, and selection protocol without proving generalizability to other projects. Stronger evidence requires replication, locked-model testing, transparent denominators, reporting of unsuccessful selections, and assessment of whether the model improves decisions relative to realistic alternatives.

A Stage-Specific Map of Credible Value, Overclaiming, and Unmet Needs

A stage-specific interpretation begins by asking what the output of an AI system actually changes. In molecular generation, low-data methods can expand focused chemical spaces and produce candidates aligned with modeled objectives even when project-specific data are limited [28]. The credible value at this stage is hypothesis generation and prioritization. The principal overclaim is to treat generated validity, novelty, or predicted activity as evidence of pharmaceutical quality. The unmet need is not merely a more powerful generator but a better connection between generation objectives and experimentally measurable properties, synthetic feasibility, multiparameter optimization, uncertainty, and downstream developability.

Hit finding occupies a stronger evidence position when models are trained on high-throughput experimental data and their predictions are tested prospectively. Machine learning applied to DNA-encoded-library selections has demonstrated that large experimental datasets can be used to prioritize compounds for defined protein targets and subsequent testing [29]. This represents credible task-level value because the model contributes directly to the selection of candidates examined experimentally. The boundary is that the workflow depends on specialized data-generation platforms and does not establish that selected hits will become optimized leads or clinical candidates. The next validation need concerns replication across targets, comparison with alternative selection strategies, and longitudinal assessment of what happens after initial hit confirmation.

Blinded or hidden-data challenges provide an intermediate evidence tier between familiar public benchmarks and full prospective deployment. A crowdsourced kinase-inhibitor prediction challenge evaluated methods against unpublished bioactivity data across comparatively underexplored target space [30]. This design reduces opportunities for direct benchmark reuse and can reveal how methods behave on data not available during development. It still evaluates a bounded prediction problem, however, and may not reproduce the compound-selection pressures, assay constraints, or iterative learning of an active discovery program. Hidden evaluation strengthens evidence for generalization but does not substitute for prospective experimental decision testing.

Clinical-pipeline evidence occupies the highest and least mature tier in the proposed map. Early analyses of AI-associated drug candidates entering clinical trials provide useful observations about emerging pipelines, but they remain vulnerable to small and changing samples, selective inclusion, historical comparison, and uncertain attribution of outcomes to AI [31]. A clinical candidate reflects target selection, medicinal chemistry, formulation, pharmacology, toxicology, manufacturing, trial design, financing, and organizational judgment in addition to computational tools. The defensible synthesis is therefore asymmetric: evidence is comparatively strong that AI changes selected search and prioritization activities, moderate where predictions are externally or prospectively confirmed, and limited where claims concern improved clinical or portfolio outcomes. **Figure 1** presents the ai value-and-limit translation map, showing how the article's key components and boundaries are connected within a stage-specific map of credible value, overclaiming, and unmet needs.

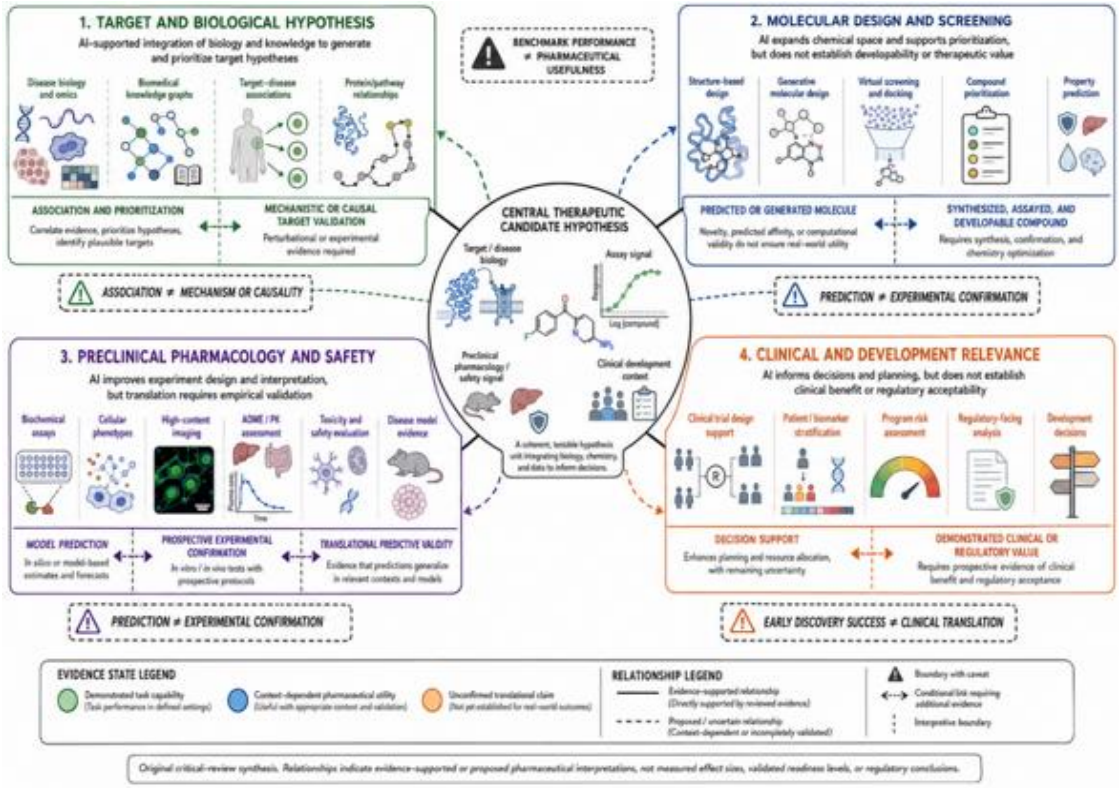

Figure 1. AI Value-and-Limit Translation Map.

The figure is an original conceptual synthesis that organizes the article's central contribution across review scope, analytical questions, and critical evaluation approach, review scope, analytical questions, critical evaluation approach, aI contributions

to target discovery, molecular design, and screening, AI contributions to target discovery. 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.

Priorities for Responsible and Evidence-Proportionate Adoption

The first priority is to align data, model development, and evaluation with a precisely defined pharmaceutical use. Best-practice recommendations for machine learning in chemistry emphasize transparent datasets, meaningful baselines, appropriate evaluation procedures, and complete reporting of methodological choices [32]. These practices should be extended from publication quality to decision quality. Before a model is adopted, users should specify what decision it supports, what alternative it replaces or supplements, what errors are consequential, what chemical or biological domain it covers, and when it should abstain. A favorable benchmark result should permit only the claim demonstrated by that benchmark unless stronger evidence is available.

The second priority is reproducibility through artifacts and documentation sufficient for independent scrutiny. Reproducibility standards for life-science machine learning call for accessible data and code where feasible, detailed reporting, and verification that analyses can be reconstructed [33]. Pharmaceutical settings introduce legitimate restrictions involving proprietary chemistry, patient information, and confidential development data. Such restrictions do not remove the need for reproducibility; they change how it must be achieved. Versioned datasets, model specifications, audit trails, locked evaluation protocols, controlled internal replication, and qualified external review can provide evidence when public release is impossible. Reproducibility should also cover data preprocessing, endpoint construction, exclusions, hyperparameter selection, and failed analyses rather than only the final model.

The third priority is context-of-use-based credibility. Scientific and regulatory evaluation of mechanistic in silico models has emphasized that credibility requirements should be proportional to the model's intended role and the consequence of its use [34]. The same principle is applicable, with appropriate caution, to AI-enabled pharmaceutical models. A system that proposes compounds for exploratory testing may require evidence of chemical validity, domain coverage, and enrichment relative to a baseline. A system used to exclude compounds from safety testing, guide clinical development, or support regulatory evidence requires substantially stronger validation, uncertainty characterization, change control, and monitoring. Evidence proportionality prevents both extremes: rejecting useful low-risk tools because they lack clinical validation and accepting high-consequence tools on the basis of retrospective computational performance.

The fourth priority is structured validation reporting that makes evidence boundaries visible. DOME recommendations organize supervised biological machine learning around data, optimization, model, and evaluation domains [35]. Applied to drug discovery, such reporting should specify data provenance, chemical and biological splits, model-selection procedures, external evaluation, applicability limits, uncertainty behavior, prospective testing, and the relationship between the predicted endpoint and the pharmaceutical decision. Reporting guidance cannot guarantee scientific validity, but it can expose where evidence is missing and reduce the rhetorical movement from performance to readiness. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop priorities for responsible and evidence-proportionate adoption within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Where Artificial Intelligence Changes Drug Discovery and Where It Does Not A

Approach or Application	Data and Task Context	Evaluation Practice	Evidence Maturity	Translational Limitation	Review Interpretation
Target Prioritization and Biomedical Graph Learning	Curated knowledge graphs, omics data, literature-derived relations, and interaction networks	External biological datasets, perturbational evidence, and comparison with non-graph baselines	Emerging task utility	Associations may reflect database structure, research concentration, or non-causal relationships	Appropriate for hypothesis prioritization; insufficient alone for causal target validation
Generative Molecular Design	Public and project-specific chemical structures, activity labels, and modeled property objectives	Chemical-validity checks, similarity analysis, prospective synthesis, assays, and multiparameter follow-up	Strong computational capability; selected prospective confirmation	Objective functions incompletely represent synthesizability, exposure, safety, formulation, and efficacy	Credible as a candidate-generation tool when followed by medicinal-chemistry and experimental evaluation
Virtual Screening and Property Prediction	Molecular structures, protein information, assay labels, and historical screening data	Scaffold or temporal splits, hidden external data, realistic baselines, applicability-domain analysis, and prospective assays	Variable by endpoint and validation setting	Leakage, chemical similarity, assay artifacts, and distribution shift can inflate apparent performance	Useful for ranking experiments; not a substitute for experimental confirmation
High-Content Imaging and Phenotypic Modeling	Cellular images, perturbation profiles, morphology, and assay metadata	Cross-batch and cross-laboratory validation, locked preprocessing, biological replication, and orthogonal assays	Emerging preclinical utility	Image representations may not transfer across protocols and do not independently establish mechanism	Valuable for scalable phenotype organization and screening when biological interpretation is separately validated

Toxicity and Safety Prediction	Heterogeneous preclinical, chemical, biological, and adverse-event endpoints	Endpoint-specific external validation, calibration, rare-event analysis, and species or population transfer assessment	Uneven and endpoint-dependent	Safety outcomes are heterogeneous, sparse, and sensitive to biological context	May support prioritization and risk identification; cannot serve as a universal safety determination
Clinical-Trial And Portfolio Support	Trial registries, protocol variables, target evidence, development histories, and multimodal clinical data	Temporal and sponsor-external validation, prospectively locked models, decision-impact assessment, and transparent outcome definitions	Early translational evidence	Confounding, historical sponsor behavior, changing standards of care, and uncertain causal attribution	May assist risk stratification; does not establish drug efficacy or an AI-caused improvement in success
Regulatory-Facing Ai Use	Model outputs included in defined drug-development analyses or submissions	Context-of-use documentation, traceability, verification, validation, uncertainty assessment, and change control	Practical use established; acceptance remains case-specific	Use in a submission is not equivalent to endorsement of a general model or method	Evidence should be judged according to the role of the model in the specific regulatory question
Routine Pharmaceutical Deployment	Integrated models used repeatedly within operational discovery or development workflows	Ongoing performance monitoring, drift detection, incident review, revalidation, governance, and human oversight	Limited public evidence	Proprietary environments obscure denominators, failures, and comparative impact	Routine use should remain bounded by documented evidence, monitoring, and explicit decision authority

Table note: Evidence-maturity descriptions are qualitative and context-dependent. They do not constitute validated readiness grades, clinical recommendations, or regulatory classifications.

Figure 2 presents the evidence, validation, and translation boundary observatory for where artificial intelligence changes drug discovery and, showing how the article's key components and boundaries are connected within priorities for responsible and evidence-proportionate adoption.

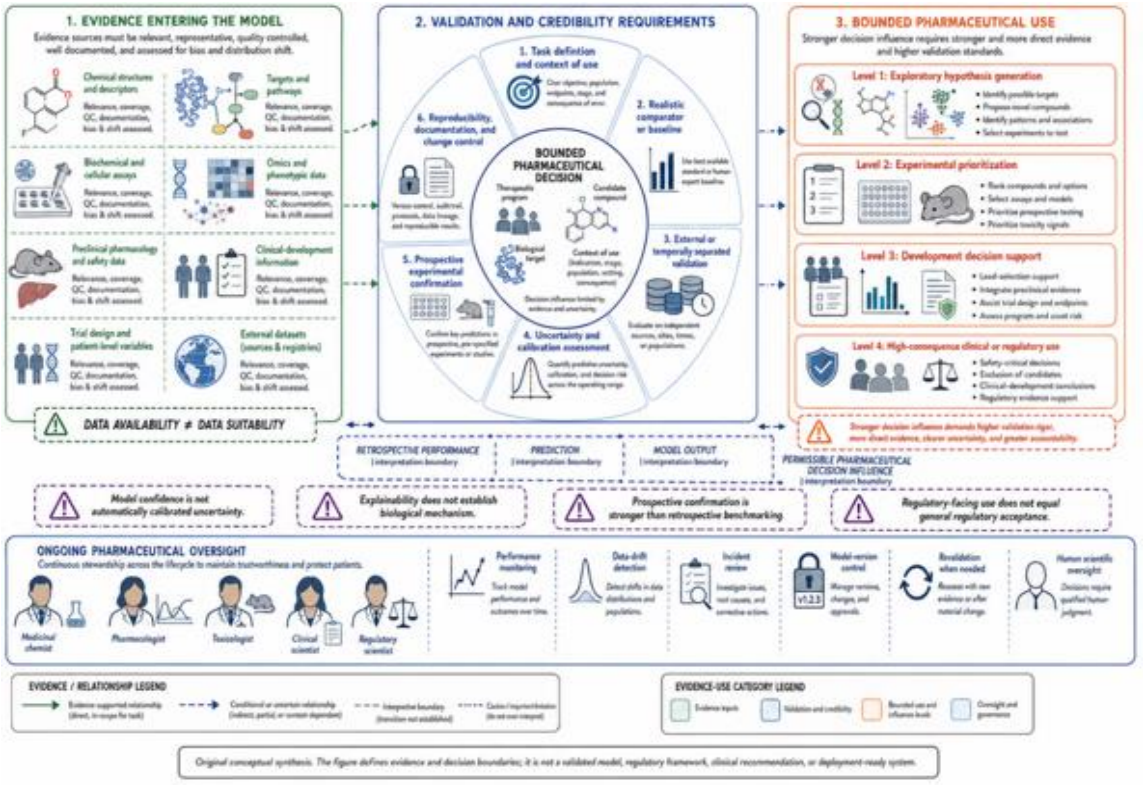

Figure 2. Evidence, Validation, and Translation Boundaries.

The figure is an original conceptual synthesis that shows how claims move from conceptual plausibility through evaluation, uncertainty assessment, and bounded pharmaceutical use. 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.

Conclusion

Artificial intelligence changes drug discovery most convincingly where it expands search, structures complex data, prioritizes candidates, or connects computation to a defined experimental decision. Its contribution is less defensible when benchmark

performance is treated as evidence of mechanism, developability, clinical benefit, regulatory acceptability, or improved portfolio success. The central contribution of this critical review is a stage-specific value-and-limit interpretation that separates computational capability, plausible utility, external validation, prospective experimental confirmation, translational evidence, and routine pharmaceutical readiness. This interpretation shows that AI value is neither absent nor uniform. It is strongest in bounded tasks with suitable data, realistic evaluation, and rapid experimental feedback, and progressively more conditional as claims move toward causal biology, preclinical translation, clinical outcomes, and high-consequence decisions. Responsible adoption should therefore be based on the context of use, the strength and directness of validation, the uncertainty represented and omitted, and the permissible influence of the model on human judgment. AI should be treated as a means of improving particular scientific and operational decisions rather than as an independent guarantee of pharmaceutical success.

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. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
3. 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
4. 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
5. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26(1):80-93. doi:10.1016/j.drudis.2020.10.010
6. Snyder H. Literature review as a research methodology: An overview and guidelines. J Bus Res. 2019;104:333-9. doi:10.1016/j.jbusres.2019.07.039
7. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372. doi:10.1136/bmj.n71
8. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358. doi:10.1136/bmj.j4008
9. Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform. 2021;22(6). doi:10.1093/bib/bbab159
10. 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
11. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702.e13. doi:10.1016/j.cell.2020.01.021
12. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76. doi:10.1021/acscentsci.7b00572
13. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023;616(7958):673-85. doi:10.1038/s41586-023-05905-z
14. Basile AO, Yahi A, Tatonetti NP. Artificial intelligence for drug toxicity and safety. Trends Pharmacol Sci. 2019;40(9):624-35. doi:10.1016/j.tips.2019.07.005
15. Carreras-Puigvert J, Spjuth O. Artificial intelligence for high content imaging in drug discovery. Curr Opin Struct Biol. 2024;87:102842. doi:10.1016/j.sbi.2024.102842
16. Harrer S, Shah P, Antony B, Hu J. Artificial intelligence for clinical trial design. Trends Pharmacol Sci. 2019;40(8):577-91. doi:10.1016/j.tips.2019.05.005
17. Liu Q, Huang R, Hsieh J, Zhu H, Tiwari M, Liu G, et al. Landscape analysis of the application of artificial intelligence and machine learning in regulatory submissions for drug development from 2016 to 2021. Clin Pharmacol Ther. 2023;113(4):771-4. doi:10.1002/cpt.2668

18. Aliper A, Kudrin R, Polykovskiy D, Kamya P, Tutubalina E, Chen S, et al. Prediction of clinical trials outcomes based on target choice and clinical trial design with multi-modal artificial intelligence. Clin Pharmacol Ther. 2023;114(5):972-80. doi:10.1002/cpt.3008
19. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A
20. 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
21. 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
22. Thakkar A, Kogej T, Reymond JL, Engkvist O, Bjerrum EJ. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. Chem Sci. 2020;11(1):154-68. doi:10.1039/C9SC04944D
23. Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. Nat Rev Chem. 2022;6(6):428-42. doi:10.1038/s41570-022-00391-9
24. 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
25. 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
26. Scannell JW, Bosley J, Hickman JA, Dawson GR, Truebel H, Ferreira GS, et al. Predictive validity in drug discovery: What it is, why it matters and how to improve it. Nat Rev Drug Discov. 2022;21(12):915-31. doi:10.1038/s41573-022-00552-x
27. Giblin KA, Hughes SJ, Boyd H, Hansson P, Bender A. Prospectively validated proteochemometric models for the prediction of small-molecule binding to bromodomain proteins. J Chem Inf Model. 2018;58(9):1870-88. doi:10.1021/acs.jcim.8b00400
28. Moret M, Friedrich L, Grisoni F, Merk D, Schneider G. Generative molecular design in low data regimes. Nat Mach Intell. 2020;2(3):171-80. doi:10.1038/s42256-020-0160-y
29. McCloskey K, Sigel EA, Kearnes S, Xue L, Tian X, Moccia D, et al. Machine learning on DNA-encoded libraries: A new paradigm for hit finding. J Med Chem. 2020;63(16):8857-66. doi:10.1021/acs.jmedchem.0c00452
30. Cichońska A, Ravikumar B, Allaway RJ, Wan F, Park S, Isayev O, et al. Crowdsourced mapping of unexplored target space of kinase inhibitors. Nat Commun. 2021;12(1):3307. doi:10.1038/s41467-021-23165-1
31. Jayatunga MKP, Ayers M, Bruens L, Jayanth D, Meier C. How successful are AI-discovered drugs in clinical trials? A first analysis and emerging lessons. Drug Discov Today. 2024;29(6):104009. doi:10.1016/j.drudis.2024.104009
32. Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8. doi:10.1038/s41557-021-00716-z
33. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-5. doi:10.1038/s41592-021-01256-7
34. 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
35. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, 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