Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

FROM ALGORITHMIC PREDICTION TO THERAPEUTIC EVIDENCE: ARTIFICIAL INTELLIGENCE IN DRUG DISCOVERY, TRANSLATION, AND RESPONSIBLE GOVERNANCE

Paola Carolina Rainer da Silva
Faculdade Presidente Antônio Carlos, Brazil
Corresponding author: rainerpaola016@gmail.com
Orcid ID: https://orcid.org/0009-0003-1526-3989

Abstract

Artificial intelligence (AI) now supports target identification, molecular representation, virtual screening, structure prediction, de novo design, synthesis planning, drug repurposing, and prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET). These capabilities can expand the chemical and biological hypotheses evaluated before costly experiments, but computational novelty is not equivalent to a safe or effective medicine. This structured critical narrative review examines AI as part of an iterative discovery system linking data, algorithms, medicinal chemistry, experimental biology, pharmacology, manufacturing, and clinical development. The synthesis emphasizes evidence quality, applicability domain, prospective validation, reproducibility, and regulatory credibility. A detailed case analysis of halicin shows both the legitimate achievement and the boundary of AI-prioritized discovery: a graph neural network selected a structurally unusual antibacterial candidate, followed by laboratory and animal validation, yet the finding did not establish clinical efficacy or approval. Additional analysis covers AlphaFold-enabled structural hypothesis generation, AI-designed molecules, multi-objective optimization, ADMET modeling, automated laboratories, clinical-development support, and AI-enabled target discovery. The article proposes five author-defined evidence levels and an author-synthesized stage-gated governance framework. Major limitations include biased or noisy data, distribution shift, shortcut learning, inadequate uncertainty estimation, proprietary opacity, environmental cost, intellectual-property ambiguity, and unequal access to data and computation. AI can accelerate learning and improve prioritization, but therapeutic value requires progressively stronger experimental, translational, clinical, and, where applicable, regulatory evidence.

Keywords: artificial intelligence; drug discovery; machine learning; halicin; virtual screening; generative chemistry; ADMET; translational science; pharmaceutical research; AI governance

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

1. Introduction

Drug discovery converts uncertain biological hypotheses into interventions that must satisfy efficacy, safety, pharmacokinetics, manufacturability, and clinical need. Attrition is high because success at one stage does not guarantee success at the next: target modulation may not alter disease, promising potency may coexist with toxicity, and animal findings may not translate to humans. Consequently, the central problem is not simply generating more molecules but learning efficiently under biological uncertainty (Paul et al., 2010; Scannell et al., 2012).

AI has attracted attention because pharmaceutical research produces large and heterogeneous datasets, including sequences, structures, images, omics profiles, assays, electronic records, and scientific text. Machine learning can rank hypotheses and compounds within these data, while deep learning can learn representations from less engineered inputs. Generative models can propose structures rather than only score existing ones (Vamathevan et al., 2019; Schneider et al., 2020).

Claims of acceleration require precise interpretation. A model may reduce the time required to screen a virtual library or propose a lead, but total development time includes synthesis, experimental replication, toxicology, formulation, manufacturing, regulatory review, and clinical trials. Faster computation can also create more downstream work if candidates are poorly constrained. The relevant outcome is improved decision quality across the pipeline, not the number of predictions produced (Bender & Cortés-Ciriano, 2021a).

The field contains both important scientific achievements and promotional overstatement. Retrospective benchmark success is often presented as evidence of practical transformation, although random train-test splits can place close analogues on both sides and inflate performance. Prospective validation, comparison with strong baselines, chemical-series separation, and disclosure of failed predictions are needed to establish real utility (Bender & Cortés-Ciriano, 2021a, 2021b).

This review asks where AI creates defensible value in drug discovery and which evidence is required before a computational output can influence development. It covers methods, pipeline applications, the halicin case, translational examples, risks, governance, and research priorities. The review therefore examines not only what AI systems can predict, but also how their outputs become credible evidence through experimental testing, pharmaceutical judgment, documentation, and accountable decision-making.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

2. Methodology

2.1 Review design

A critical narrative literature review was selected because AI-enabled drug discovery spans computer science, cheminformatics, structural biology, pharmacology, medicinal chemistry, clinical development, and regulation, while available studies use fundamentally different tasks, datasets, comparators, and endpoints. The purpose was explanatory rather than exhaustive: to identify where AI contributes defensible value, compare evidence maturity across pipeline stages, and derive governance requirements that connect computational outputs to experimental and therapeutic decisions. The manuscript is therefore not presented as a systematic review or meta-analysis (Snyder, 2019; Whittemore & Knafl, 2005).

2.2 Information sources and search strategy

One reviewer conducted the searches, screening, extraction, appraisal, and synthesis. The final search update was performed on September 6, 2026. PubMed/MEDLINE, IEEE Xplore, Google Scholar, Crossref metadata, and the websites of the U.S. Food and Drug Administration (FDA), European Medicines Agency (EMA), World Health Organization (WHO), National Institute of Standards and Technology (NIST), and United States Patent and Trademark Office (USPTO) were searched. Citation chaining from pivotal primary studies and critical reviews was used to locate foundational papers, prospective validation studies, corrections, and regulatory or legal documents.

The core PubMed search was: (("artificial intelligence"[Title/Abstract] OR "machine learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "generative model*"[Title/Abstract]) AND ("drug discovery"[Title/Abstract] OR "drug development"[Title/Abstract] OR "virtual screening"[Title/Abstract] OR ADMET[Title/Abstract] OR "drug repurposing"[Title/Abstract])). A second focused search combined (halicin OR abaucin OR rentosertib OR "AlphaFold 3") with (discovery OR validation OR trial). IEEE Xplore was queried with the same concept blocks in title, abstract, and author-keyword fields. Google Scholar was used for citation chaining and exact-title verification, not as the sole basis for inclusion. Regulatory searches combined artificial intelligence, drug development, model credibility, context of use, lifecycle, and good AI practice. Searches were restricted to English; the main contemporary window was 2018-2026, with earlier landmark and methodological publications retained when necessary.

Selection-flow audit. The archived review worksheet confirms 26 analytical publications and seven methodological/regulatory sources in the core corpus, followed by three targeted supporting additions during revision. A reliable count of every initially displayed

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

record and every duplicate was not preserved across adaptive Google Scholar, Crossref, and institutional-site searches. Those unavailable numbers are therefore reported as not reconstructable rather than estimated retrospectively. Screening proceeded through title/abstract or relevance review, full-text assessment, and decision-level evidence mapping. Main exclusion categories were: no identifiable AI-supported decision; promotional or unverifiable evidence; duplicate treatment of the same claim; non-English text; and absence of decision-relevant validation or methodological contribution. The resulting architecture comprises 26 core analytical sources, seven core methodological/regulatory sources, and three targeted supporting sources.

2.3 Eligibility, source roles, and evidence mapping

An analytical source was operationally defined as a publication contributing directly to evaluation of an identifiable pipeline decision, AI method, validation level, or translational claim. Eligible analytical sources could therefore include primary experiments, benchmark studies, clinical investigations, or critical syntheses, provided their contribution could be mapped to a decision and evidence level. Priority was given to peer-reviewed primary research, prospective or experimental validation, independent benchmark assessment, and high-quality critical reviews. Methodological, reporting, legal, and regulatory documents were retained as contextual sources and were not treated as empirical proof of therapeutic benefit. Preprints, vendor marketing, company announcements without corresponding peer-reviewed evidence, Wikipedia, unverifiable claims, and articles mentioning AI without an identifiable role in a development decision were excluded.

The core evidence map contained 26 analytical publications and seven methodological, reporting, or regulatory sources. The revision added three targeted supporting sources: a contemporary ADMET benchmark/platform study, a direct computational-energy reporting framework, and current USPTO inventorship guidance. These additions support bounded domain-specific sections and do not convert the review into an exhaustive census. Sources were classified by decision point: target identification, structural modeling, hit identification, generative design, synthesis and closed-loop experimentation, ADMET and toxicity, repurposing, clinical development, or governance. For each analytical source, the reviewer recorded source type, AI contribution, validation level, principal finding, main limitation, and evidence maturity. Table 1 summarizes the analytical evidence; Supplementary Table S1 provides the audit trail.

2.4 Appraisal and synthesis

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Because the corpus included primary experiments, benchmarks, clinical research, and reviews, a single numerical risk-of-bias score would create false comparability. A structured evidence-maturity appraisal was used instead. Evidence was considered stronger when it included prespecified evaluation, chemically or temporally separated test data, strong baselines, external validation, complete reporting of tested candidates, prospective laboratory confirmation, mechanistic support, or clinical evaluation. Evidence was downgraded for random-split leakage risk, undisclosed proprietary data, reliance on surrogate endpoints, absence of experimental confirmation, selective reporting, or unsupported generalization from computational performance to therapeutic success.

The synthesis compared what decision was supported, what evidence was available at that stage, and what additional evidence was necessary for translation. Maturity categories were defined as stronger, when prospective or clinical evidence supported a bounded claim; moderate, when multiple studies and experimental validations existed but external or downstream confirmation remained incomplete; emerging, when evidence was dominated by retrospective benchmarks or early proof-of-concept; and insufficient, when claims depended mainly on promotional or non-verifiable material. These labels are narrative confidence judgments assigned using the criteria above; they are not validated therapeutic grades, do not constitute GRADE ratings, and do not rank commercial systems.

2.5 Single-reviewer procedures and limitations

Screening, source classification, extraction, and thematic interpretation were performed by one reviewer. Consequently, inter-rater reliability could not be calculated, and the evidence-maturity categories necessarily incorporate individual judgment. No kappa or percentage agreement is reported. To reduce interpretive drift, the same extraction fields and appraisal criteria were applied across the corpus; empirical, methodological, legal, corporate, and regulatory evidence were kept conceptually distinct; and the complete source-evidence matrix is supplied for audit. These procedures improve consistency but do not eliminate selection, classification, or confirmation bias.

3. Conceptual Framework: Prediction-to-Evidence Translation

This review proposes an author-synthesized, stage-gated human-in-the-loop framework informed by scientific and regulatory principles. It is an analytical integration, not a formally adopted regulatory pathway. AI capability is not treated as an outcome in itself: a model transforms data into a prediction or generated hypothesis; a multidisciplinary team interprets the output within an applicability domain; an experiment tests the decision-relevant claim; and the resulting evidence supports progression, revision, or termination.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Governance moderates each stage through provenance, validation requirements, access control, documentation, accountability, and change management.

The author proposes five evidentiary levels to prevent inferential overreach: computational plausibility, prospective prioritization, experimental validity, translational validity, and therapeutic evidence. They are analytical categories, not an established regulatory grading system. Computational plausibility indicates performance on a defined dataset. Prospective prioritization shows selection of previously untested candidates. Experimental validity requires reproducible activity with appropriate controls. Translational validity adds developability, pharmacology, and evidence in relevant biological systems. Therapeutic evidence requires adequately designed clinical studies and, where applicable, regulatory assessment. Movement between levels is not automatic because each transition introduces new failure mechanisms.

The framework also separates automation from augmentation. Automation executes a defined process with limited case-by-case judgment, whereas augmentation expands the scientist's ability to search, compare, and prioritize while preserving professional responsibility. In drug discovery, the most credible applications are usually augmentative because medicinal chemistry, biological interpretation, toxicology, clinical judgment, and regulatory accountability cannot be reduced to a single model score. Figure 1 represents this prediction-to-evidence loop.

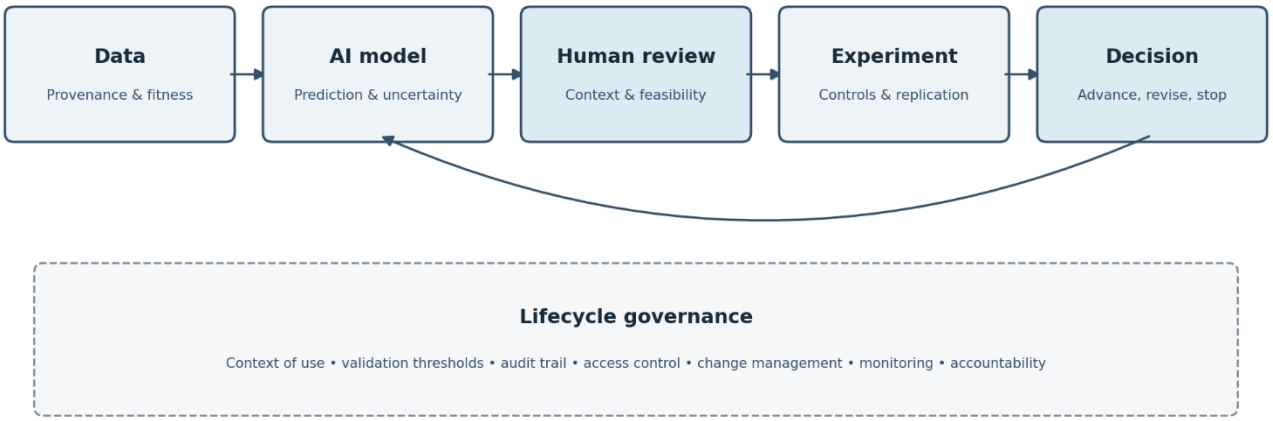


Figure 1. Prediction-to-evidence loop for AI-augmented drug discovery

Note. Conceptual framework synthesized by the author from the reviewed scientific and regulatory principles. It is not a formally endorsed regulatory pathway. AI outputs inform decisions only after domain review and decision-relevant validation; governance applies across the lifecycle.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

4. Conceptual Foundations and Molecular Representations

AI is an umbrella term; machine learning learns mappings from data, and deep learning uses multilayer neural networks to learn complex representations. Supervised learning predicts labeled outcomes such as potency or toxicity, unsupervised learning organizes unlabeled chemical or omics data, and reinforcement learning optimizes sequences of molecular actions using a reward. The method should follow the scientific question and data, not fashion (Vamathevan et al., 2019).

Molecules can be encoded as descriptors, fingerprints, graphs, strings such as SMILES, three-dimensional conformations, or combinations of these forms. Representation determines which relationships a model can learn and which invariances must be built in. Graph neural networks naturally express atoms and bonds, while sequence models exploit chemical strings; three-dimensional models address geometry but inherit uncertainty about biologically relevant conformations (Wu et al., 2018; Yang et al., 2019).

Protein and biological representations include amino-acid sequences, predicted or experimental structures, interaction networks, gene expression, and multimodal patient data. These sources operate at different scales and are often incomplete. Combining them may reveal convergent evidence, but integration can also hide batch effects and inconsistent biological contexts (Vamathevan et al., 2019).

Chemical language models learn statistical regularities from molecular strings or graphs and may be adapted to property prediction or generation. Their apparent fluency does not guarantee chemical validity, synthesizability, novelty, or target engagement. Generated molecules require sanitization, similarity checks, property filters, expert review, synthesis assessment, and experiments (Schneider et al., 2020).

An applicability domain defines where model predictions are supported by training experience. Predictions for novel scaffolds, targets, species, or assay conditions may be more uncertain than scores indicate. Distance measures, ensembles, calibration, and conformal approaches can help characterize uncertainty, but cannot substitute for new data when the system extrapolates (Bender & Cortés-Ciriano, 2021b).

5. Target Identification and Structural Biology

Target identification seeks causal leverage over disease rather than mere association. Machine learning can integrate genetics, omics, pathways, literature, and clinical observations to prioritize proteins or mechanisms. However, observational patterns may reflect downstream consequences or confounding. Genetic support, perturbational experiments, disease models, and human evidence remain necessary to strengthen causal plausibility (Vamathevan et al., 2019).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Network-based approaches model diseases and targets within interacting biological systems. They can identify modules, polypharmacology, and opportunities for repurposing, especially when a single-target model is inadequate. Network predictions are sensitive to missing and preferentially studied interactions, so the literature's historical attention can become an algorithmic prior (Zitnik et al., 2018).

AlphaFold2 substantially improved computational protein-structure prediction and demonstrated near-experimental accuracy for many targets in CASP14. The AlphaFold Protein Structure Database expanded access to predicted structures, enabling hypothesis generation where experimental structures were absent. This achievement changes structural biology workflows but does not solve every drug-design problem (Jumper et al., 2021; Varadi et al., 2022).

AlphaFold 3 extended this trajectory by modeling complexes that include proteins, nucleic acids, small molecules, ions, and modified residues. Its reported gains are important for interaction hypotheses, but they do not convert structure prediction into affinity prediction or establish that a modeled pose is pharmacologically relevant. Benchmark category, training-set overlap, model confidence, and access to reproducible workflows remain necessary parts of interpretation (Abramson et al., 2024; Akdel et al., 2022).

Predicted structures may not represent ligand-bound conformations, dynamic ensembles, post-translational states, complexes, or local uncertainty important for binding. Confidence metrics must be inspected, and experimental methods remain essential for decisive claims. Structure prediction supports prioritization and interpretation; it does not automatically establish a druggable pocket or reliable affinity (Jumper et al., 2021).

The strongest target workflows triangulate multiple sources and define a falsifiable experiment. AI should help choose what to test, not make causal language stronger than the evidence. Teams should record why a target was prioritized, which datasets contributed, what contradicting evidence exists, and which experiment could change the decision (Vamathevan et al., 2019).

6. Virtual Screening, Docking, and Hit Prioritization

Virtual screening ranks compounds before physical testing. Ligand-based models learn from known active and inactive molecules, whereas structure-based methods evaluate compatibility with a binding site. AI scoring functions can complement docking and physics-based calculations, but performance depends on pose generation, protein preparation, assay definitions, decoys, and the chemical relationship between training and test sets (Lavecchia, 2015).

Random splits often overestimate generalization because close analogues appear in both training and test data. Scaffold splits, temporal splits, and prospective screens are more

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

demanding and more relevant to discovery. Even these designs should report enrichment at the experimentally affordable top of the ranking rather than only global classification metrics (Bender & Cortés-Ciriano, 2021b).

Docking scores are approximations of a complex physical process. Protein flexibility, water, protonation, entropy, and assay context can change binding. A machine-learning rescoring model may exploit dataset artifacts rather than molecular recognition. Orthogonal calculations and experiments are required, especially when the proposed scaffold differs from known ligands (Walters & Murcko, 2020).

The value of screening depends on prospective yield: how many tested compounds become reproducible hits with acceptable controls. Publications should disclose library composition, selection rules, assay interference checks, counter-screens, and inactive predictions. Without this information, a successful molecule cannot reveal how many resources were spent or whether the model outperformed simpler selection (Bender & Cortés-Ciriano, 2021a).

Active learning can make screening iterative. The model selects informative compounds, experiments produce labels, and the next cycle balances exploitation of promising regions with exploration of uncertainty. This loop can be more valuable than a one-time ranking because it adapts to the local assay and exposes model errors early (Reker & Schneider, 2015).

7. Halicin: A Landmark in AI-Enabled Antibiotic Discovery

The discovery of halicin is a landmark because it demonstrated prospective experimental value rather than benchmark performance alone. Stokes and colleagues trained a directed-message-passing neural network on 2,335 molecules tested for inhibition of Escherichia coli growth. The representation learned molecular features from graph structure and was then used to prioritize compounds outside conventional antibiotic scaffolds (Stokes et al., 2020; Yang et al., 2019).

The model screened the Drug Repurposing Hub and highlighted SU3327, a compound previously investigated in a different therapeutic context. Researchers renamed it halicin in recognition of HAL from 2001: A Space Odyssey. Its low structural similarity to familiar antibiotics illustrated a central promise of representation learning: prioritizing activity without depending solely on explicit similarity to known classes (Stokes et al., 2020).

Halicin was not declared an antibiotic on the basis of a computer score. The team tested it experimentally, observed activity against a range of bacterial species, investigated its mechanism, evaluated resistance development under laboratory conditions, and demonstrated efficacy in murine infection models. The work therefore linked prediction to microbiology and in vivo evidence (Stokes et al., 2020).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

The study also extended screening to more than 107 million structures from the ZINC15 database. From a small set of high-ranking, structurally diverse, and practically testable predictions, the researchers identified additional antibacterial molecules. This illustrates how AI can compress a very large virtual space into a manageable experimental set, which is a prioritization achievement rather than a replacement for screening (Stokes et al., 2020).

Halicin's mechanism was associated with disruption of bacterial electrochemical gradients, providing a biological explanation distinct from many established antibiotics. Mechanistic investigation mattered because antibacterial activity can arise from nonspecific or undesirable effects. A model can identify a pattern associated with activity, but experimental science must determine how and under what conditions that activity occurs (Stokes et al., 2020).

The case should not be described as the discovery of an approved medicine. Halicin remained a preclinical candidate; animal efficacy does not establish human dosing, safety, pharmacokinetics, comparative benefit, manufacturability, or resistance behavior in clinical use. The accurate claim is that deep learning enabled the identification and prioritization of a structurally unusual antibacterial candidate that was subsequently validated in laboratory and animal studies (Stokes et al., 2020).

A published correction clarified an error in the original paper's description of the compound's prior clinical development, reinforcing the importance of provenance and careful reporting even in influential studies. Corrections are part of scientific self-correction and should be incorporated when the halicin narrative is retold (Stokes et al., 2020).

Halicin provides three durable lessons. Novelty is valuable when it is defined against appropriate chemical references; prospective validation is more persuasive than retrospective accuracy; and the model's contribution must be separated from the experimental program that establishes biological credibility. These distinctions protect a genuine accomplishment from exaggerated claims (Stokes et al., 2020; Bender & Cortés-Ciriano, 2021a).

The later discovery of abaucin provides a useful comparison. A deep-learning-guided screen prioritized a compound with selective activity against Acinetobacter baumannii, followed by experimental investigation of its activity and mechanism. Together, halicin and abaucin show that the strongest contribution is not a generic claim that AI "discovers antibiotics," but a reproducible sequence in which a model narrows a search space and microbiological experiments establish what the selected molecule actually does (Liu et al., 2023).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

8. Generative Chemistry and Multi-Objective Optimization

Generative models propose molecules using recurrent networks, variational autoencoders, generative adversarial networks, transformers, diffusion models, or graph-based construction. They can search beyond catalogued libraries and condition generation on desired properties. The challenge is that drug design is constrained by many objectives that are imperfectly predicted and often conflict (Elton et al., 2019; Schneider et al., 2020).

Optimizing a single predicted score invites reward hacking: a generator may exploit weaknesses in the property model rather than produce a better drug. Multi-objective design includes potency, selectivity, solubility, permeability, metabolic stability, safety, novelty, and synthesis, yet weighting these terms embeds scientific and commercial judgments. Pareto fronts can display trade-offs more honestly than a single composite reward (Schneider et al., 2020).

Validity and novelty are insufficient evaluation metrics. A molecule may obey valence rules and be unlike training examples while being unstable, reactive, impossible to synthesize, or biologically irrelevant. Evaluation should include medicinal-chemistry filters, synthesis planning, retrosynthetic feasibility, three-dimensional considerations, and prospective experiments against strong baselines (Walters & Murcko, 2020).

Human-in-the-loop design remains important because medicinal chemists reason about series history, assay behavior, intellectual property, formulation, and tacit synthetic knowledge. AI can generate alternatives and reveal non-obvious trade-offs, while experts reject artifacts and redefine objectives. The productive unit is the iterative team and laboratory, not the isolated generator (Schneider et al., 2020).

Closed-loop platforms connect design, robotic synthesis, assays, and model updating. MacLeod et al. (2020) provide a prospective autonomous-experimentation exemplar from thin-film materials rather than pharmaceuticals. The transfer is conceptually useful, but evidence for fully autonomous, end-to-end closed-loop drug discovery remains less mature than in selected materials domains. Pharmaceutical implementation additionally requires compound identity and purity controls, validated bioassays, pharmacological interpretation, and quality systems. Accordingly, this section supports automation of bounded cycles, not replacement of multidisciplinary drug-development judgment.

9. ADMET, Toxicity, and Developability

ADMET properties determine whether a potent compound can reach the relevant tissue at a tolerable exposure. Models predict solubility, permeability, clearance, metabolism, transporter interactions, cardiac liabilities, and toxicological endpoints. Earlier prediction

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

can remove weak candidates before expensive studies, but labels are heterogeneous and often protocol-dependent (Yang et al., 2019).

Toxicity is especially difficult because mechanisms vary by dose, duration, species, organ, metabolites, and patient susceptibility. A negative model output is not proof of safety. Regulatory toxicology requires validated assays and context-specific evidence; computational models are most useful for prioritizing tests, interpreting mechanisms, and supporting weight-of-evidence assessments (Vamathevan et al., 2019).

Calibration and uncertainty are critical when predictions determine whether a compound advances. Models should state the domain in which they were validated and identify out-of-domain chemistry. Decision thresholds should reflect the asymmetric cost of missing a serious liability versus discarding a potentially useful series (Bender & Cortés-Ciriano, 2021b).

Multi-parameter optimization should begin before potency is maximized. Late discovery of insolubility, metabolic instability, or selectivity problems creates costly redesign. Integrated models can make trade-offs visible, but correlated prediction errors may create false confidence when multiple endpoints share biased training data (Schneider et al., 2020).

A contemporary primary example, ADMET-AI, trained graph-neural-network models across 41 datasets and reported comparative performance on the Therapeutics Data Commons ADMET benchmark, open code, and archived models (Swanson et al., 2024). This strengthens evidence for scalable computational prioritization across multiple endpoints, but it remains chiefly benchmark validation rather than prospective proof that use of the system reduces clinical attrition. Endpoint-specific external, temporal, and prospective validation therefore remains necessary, particularly for rare toxicities and shifted chemistry.

Pharmacists and pharmaceutical scientists contribute expertise in metabolism, formulation, interactions, and clinically meaningful exposure. Their participation helps connect computational endpoints to real dosage forms, administration routes, patient populations, and monitoring requirements rather than treating developability as a purely molecular score (Vamathevan et al., 2019).

10. Drug Repurposing and Knowledge Integration

Drug repurposing searches for new uses of compounds with existing biological or clinical information. AI can combine targets, pathways, signatures, networks, literature, adverse-event data, and health records to rank hypotheses. Existing knowledge may shorten some early work, but a new indication still requires mechanistic, dosing, safety, and clinical evidence (Pushpakom et al., 2019).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Knowledge graphs represent entities and relations across drugs, genes, diseases, and phenotypes. Link-prediction models can expose plausible connections that are difficult to inspect manually. Their outputs inherit missing, disputed, and duplicated facts, and highly studied entities may dominate. Evidence provenance should remain visible for every predicted relation (Zitnik et al., 2018).

The COVID-19 period demonstrated both the speed and the danger of computational repurposing. Numerous candidates were proposed from signatures, networks, docking, and text mining, but biological plausibility did not ensure clinical benefit. Controlled trials remained decisive, illustrating why urgent need cannot lower the distinction between hypothesis generation and therapeutic evidence (Pushpakom et al., 2019).

Real-world data can support repurposing through observational associations, yet treatment selection, disease severity, access, and documentation create confounding. Causal-inference design and prospective confirmation are needed before associations guide care. Models that predict outcomes from health records should not be described as discovering efficacy without these safeguards (Vamathevan et al., 2019).

Repurposing is strongest when computational evidence converges with pharmacology, achievable exposure, disease biology, and a feasible trial. The output should be a ranked and testable rationale, not a list of attractive connections. Negative experiments should update both the hypothesis and the model (Pushpakom et al., 2019).

11. Translation to Clinical Development

AI-assisted programs are beginning to enter clinical development, but terminology should identify the actual contribution. In this review, AI-enabled target discovery means that AI supports target nomination; AI-prioritized means that a model ranks existing candidates; AI-designed or AI-generated means that computational generation materially contributes to molecular design; and AI-supported clinical development means that AI assists a clinical-development decision. The broader term AI-assisted is used when several human and computational stages are combined. Trial entry is a milestone rather than proof of therapeutic success, and the phrase ‘AI-discovered drug’ is avoided when AI performed only one stage.

Clinical-development applications include eligibility matching, site selection, recruitment forecasting, endpoint modeling, dose selection, safety surveillance, and external-control research. Each application changes a decision and requires validation in the intended context. Operational efficiency cannot compensate for biased eligibility or unreliable endpoint ascertainment (Zhang et al., 2025).

Biomarker-driven enrichment can increase statistical power by selecting patients more likely to respond, but it may narrow generalizability and access. Biomarkers require

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

analytical and clinical validation, and subgroup hypotheses should be protected from overfitting. Trial design must preserve interpretable inference rather than adapt until a favorable pattern appears (FDA, 2025).

Governance assurance should be proportionate to context of use. Exploratory research models prioritize hypotheses for further testing; development-decision models influence candidate progression, formulation, or trial planning; clinical-decision models may affect participant-level or trial-level decisions; and regulatory-submission models produce information intended to support formal conclusions about product safety, effectiveness, or quality. In January 2025, FDA issued Draft Level 1 guidance—nonbinding and not for implementation—on the last of these contexts. The document provides a risk-based credibility framework for AI model outputs used in regulatory decision-making and explicitly does not address discovery-stage models or operational efficiencies that do not affect safety, quality, or study reliability (FDA, 2025).

The ultimate test of AI-enabled discovery is not whether a model proposed a compound quickly, but whether the resulting program produces reproducible evidence of patient benefit with acceptable risk. Many years may separate an impressive computational result from that determination. Responsible reporting maintains this temporal and evidentiary perspective (Zhang et al., 2025).

Rentosertib provides a recent translational example with peer-reviewed clinical evidence. The program combined AI-enabled target identification and AI-designed molecular generation with conventional synthesis, preclinical development, clinical operations, and human oversight. A multicenter randomized phase 2a trial in idiopathic pulmonary fibrosis included 71 participants and reported safety and tolerability alongside exploratory pharmacokinetic, biomarker, and clinical outcomes. This supports a bounded claim of clinical translation; it does not establish approval, definitive comparative efficacy, or a general success rate for AI-assisted programs. Larger confirmatory trials and a clearly reported denominator of all comparable programs are required before making comparative productivity claims (Xu et al., 2025).

12. Data Quality, Reproducibility, and Benchmark Failure

Biochemical data combine different protocols, instruments, species, cell lines, units, and curation practices. “Inactive” may mean tested and negative, below a threshold, or simply unreported. Models can learn laboratory identity or publication patterns instead of biology. Harmonization, metadata, replicate analysis, and assay-aware splits are therefore foundational (Bender & Cortés-Ciriano, 2021b).

Data leakage occurs when information from the test set influences training through duplicates, close analogues, preprocessing, or target overlap. Leakage can produce

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

spectacular but meaningless performance. Benchmark authors should document deduplication, split logic, hyperparameter selection, and whether chemical or biological similarity crosses partitions (Wu et al., 2018).

Strong baselines matter. Random forests, similarity search, docking, simple descriptors, and domain heuristics can outperform complex models in limited datasets. A deep model is useful when it improves a relevant decision enough to justify computation, maintenance, and opacity, not because it is newer (Bender & Cortés-Ciriano, 2021a).

Reproducibility requires code, data or permissible equivalents, environment, seeds, preprocessing, hyperparameters, and evaluation rules. Proprietary programs may not disclose all assets, but decision makers still need auditable documentation and independent testing. A result tied to an undocumented system version cannot support durable scientific claims (Wilkinson et al., 2016).

Negative results are valuable because they reveal limits and improve active-learning cycles. Publication systems favor successful candidates, making hit rates and cost difficult to estimate. Prospective studies should report the complete tested set and prespecified selection procedure whenever confidentiality allows (Bender & Cortés-Ciriano, 2021a).

13. Ethics, Intellectual Property, Security, and Equity

AI-enabled discovery raises questions about training-data rights, generated structures, patentability, inventorship, and freedom to operate. Legal treatment varies by jurisdiction and cannot be generalized globally. In the United States, revised USPTO guidance states that the ordinary inventorship standard applies to AI-assisted inventions and that only natural persons can be named inventors; the analysis turns on the human contribution to the claimed invention (USPTO, 2025). This does not resolve ownership, disclosure, or patentability questions in other jurisdictions. Specialized legal review and prior-art searching should therefore begin early because a computationally novel molecule may still be unpatentable or commercially unusable.

Sensitive human omics and clinical data require purpose limitation, governance, security, and respect for participants. De-identification may not eliminate re-identification risk in genomic datasets. Data access should be proportionate, logged, and governed by agreements that address secondary use and model training (WHO, 2021).

Cybersecurity is relevant because models, chemical databases, laboratory automation, and cloud infrastructure are connected assets. Poisoned datasets, manipulated assay results, stolen models, or compromised robotic workflows can corrupt scientific decisions. Security controls and incident response should be integrated into laboratory quality systems (NIST, 2023).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Unequal access to curated data, specialized talent, compute, and automated laboratories may concentrate benefits in wealthy firms and institutions. Models trained on populations and diseases that attract investment may neglect conditions important in low-resource settings. Funding and infrastructure policy should support locally relevant datasets and public-interest research (WHO, 2021).

Environmental costs arise from model training, hyperparameter search, repeated inference, computation-intensive simulation, and automated experimentation. Direct reporting frameworks show that energy use and carbon emissions depend on hardware, runtime, utilization, electricity source, and location; performance claims alone therefore provide an incomplete account of efficiency (Henderson et al., 2020). Authors should report compute and energy where feasible, reuse trained representations, compare smaller baselines, and reserve high-cost computation for decision-relevant gains. These measures support transparent accounting without assuming that every AI application has the same environmental burden.

14. Stage-Gated Governance and Implementation

A governance program begins by classifying the context of use: exploratory research, development decision, clinical decision, or regulatory submission. Assurance increases with the consequence of error. A model that ranks molecules for laboratory testing requires fit-for-purpose technical validation and experimental confirmation; a model affecting clinical evidence or a submission additionally requires regulator-ready credibility evidence, traceability, change control, and documented risk mitigation. Prohibited uses and escalation routes should be written before deployment (NIST, 2023; FDA, 2025).

Data governance should record source, license, assay conditions, curation, transformations, exclusions, and version. Datasheets and model cards can summarize intended use and limitations, but they must be connected to laboratory records and decisions. Provenance enables correction when a source or label changes (Wilkinson et al., 2016).

Validation should progress from internal testing to external and temporal datasets, then to prospective experimental evaluation. Metrics must match the decision: early enrichment and hit yield for screening, calibration for risk, and multi-objective success for lead optimization. Prespecified thresholds reduce selective interpretation (Bender & Cortés-Ciriano, 2021b).

Human review should be substantive. Medicinal chemists, biologists, pharmacologists, toxicologists, data scientists, pharmacists, clinicians, and regulatory specialists contribute different failure detection. Reviewers need access to underlying evidence, uncertainty, and alternatives rather than a single score (Vamathevan et al., 2019).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Change management covers new training data, code, model weights, prompts, descriptors, assay protocols, and infrastructure. Material changes require regression tests and sometimes renewed prospective validation. Performance should be monitored for drift, and teams must be able to suspend a model without stopping the scientific program (NIST, 2023).

The joint FDA-EMA Guiding Principles of Good AI Practice in Drug Development, published in January 2026, articulate lifecycle-oriented, human-centric principles concerning context of use, risk-proportionate evidence, data governance, multidisciplinary expertise, performance assessment, documentation, and monitoring. They are guiding principles rather than binding approval criteria. They complement rather than replace FDA's January 2025 draft credibility framework, which remains Draft Level 1 guidance and is specific to AI-generated information supporting regulatory decisions. Neither document independently validates a discovery-stage model or establishes therapeutic efficacy (FDA, 2025; FDA & EMA, 2026).

15. Evidence-Maturity Synthesis

The evidence map indicates that maturity is not uniform across the pipeline. Protein-structure prediction has exceptionally strong benchmark achievements, but drug-design relevance still depends on conformational state and experimental context. Machine-learning-guided antibacterial screening has persuasive prospective exemplars, although the number of successes remains small relative to the diversity of pathogens, targets, and chemical spaces. Generative chemistry has advanced from molecular proposal to clinical evaluation in at least one prominent program, yet evidence about comparative success rates, total costs, and reproducibility across organizations remains limited.

ADMET and toxicity prediction occupy an intermediate and endpoint-dependent position. Open benchmarks and tools such as ADMET-AI provide reproducible comparative evidence across multiple endpoints, but retrospective benchmark leadership is not prospective therapeutic validation (Swanson et al., 2024). Assay heterogeneity, class imbalance, species differences, distribution shift, and scarce reliable negative labels constrain generalization. Repurposing models generate ranked hypotheses, but observational associations and network proximity remain vulnerable to confounding and historical research bias. Clinical-development uses also have distinct evidentiary targets: recruitment efficiency does not establish endpoint validity, and a predictive biomarker does not establish treatment benefit.

Across domains, three conditions distinguish credible augmentation from promotional acceleration. First, the model must address a decision whose success can be measured. Second, evaluation must reproduce the novelty, shift, and resource constraints of

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

intended use. Third, downstream experiments must be reported with enough completeness to reveal false positives, failures, and verification costs. Table 2 summarizes the resulting maturity assessment, while Table 3 specifies stage-gated governance responsibilities. Recent pipeline-wide reviews reach a similar conclusion: current utility is real but uneven, and clinical or economic claims must be separated from computational acceleration (Fu & Chen, 2025; Niazi & Mariam, 2025).

16. Discussion

AI's most reliable contribution is prioritization under constraint. It can reduce a large search space, integrate evidence, and propose experiments, thereby increasing the amount learned per cycle. It does not remove biological uncertainty; it redistributes attention and resources. Claims should therefore be framed in terms of decisions improved and experiments enabled (Vamathevan et al., 2019).

Halicin exemplifies this balanced interpretation. The neural network enabled selection of a structurally distinctive antibacterial candidate, while laboratory and animal experiments established the reported biological findings. The model and experiments were complementary. Calling the result an approved AI-discovered antibiotic would be false; calling it merely a computational prediction would understate the prospective validation (Stokes et al., 2020).

Generative chemistry expands hypothesis production, but drug discovery is usually limited by reliable evaluation. As generators improve, assay capacity, synthesis quality, mechanistic understanding, and data governance become more important. The field should optimize closed learning loops rather than maximize molecular novelty in isolation (Schneider et al., 2020).

Evidence remains uneven across the pipeline. Structure prediction and selected screening applications have clear scientific achievements, while broad claims about reduced cost, time, or clinical success are harder to establish. Company portfolios are informative but subject to selection and commercial reporting. Independent longitudinal evidence is needed (Bender & Cortés-Ciriano, 2021a; Zhang et al., 2025).

The appropriate institutional strategy is selective adoption with rigorous measurement. Teams should begin with high-value, testable tasks; compare against strong existing methods; integrate domain experts; and expand only after prospective evidence. This approach supports innovation while protecting resources and scientific credibility (FDA, 2025; NIST, 2023).

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

17. Research Priorities

Future research should emphasize prospective head-to-head studies that disclose the full nominated and tested sets. Outcomes should include experimental hit rate, novelty, selectivity, developability, cycle time, cost, and downstream progression. Shared benchmarks remain useful, but prospective evidence should carry greater weight (Bender & Cortés-Ciriano, 2021a).

Benchmark design should use scaffold, temporal, target, and laboratory shifts that resemble future use. Evaluation should report uncertainty and performance for rare or difficult chemistry. Community challenges should preserve hidden test sets and refresh them to reduce benchmark overfitting (Wu et al., 2018).

Multimodal models need research on causal validity and missingness. Combining structure, omics, images, and text can improve prediction while creating more routes for shortcuts. Ablation studies, perturbational validation, and counterfactual testing can help determine what evidence the model actually uses (Vamathevan et al., 2019).

Research should also examine organizational outcomes: how scientists interpret models, when they override them, whether automation narrows exploration, and how incentives affect reporting. Human factors can determine whether an accurate tool improves science or creates confirmation bias (NIST, 2023).

18. Limitations

This review has limitations. It is a structured, explanatory critical narrative review rather than a systematic census, and each large pipeline domain is represented selectively by decision-relevant exemplars. The initial record and duplicate counts were not fully preserved across adaptive search sources and cannot be reconstructed without false precision. English-language restriction may omit relevant work. One reviewer performed screening, extraction, appraisal, and synthesis; inter-rater reliability is therefore unavailable. The framework combines heterogeneous evidence and its maturity labels are narrative judgments rather than validated grades. Rapid technological and regulatory change may alter conclusions after the September 6, 2026 search date. Proprietary evidence creates verification asymmetry: peer-reviewed prospective validation, corporate pipeline disclosure, company press releases, and independently reproducible evidence must not be treated as equivalent. Fully autonomous pharmaceutical closed-loop evidence remains limited, and the ADMET section cannot represent every endpoint-specific literature. These boundaries constrain generalization but reinforce the article's central demand for claim-specific evidence.

All screening, extraction, and classification were performed by one reviewer. Inter-rater reliability was therefore unavailable, and the evidence categories may reflect individual

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

selection or interpretation. The English-language restriction may have excluded regional evidence. Google Scholar was used for citation chaining, which improves coverage but is less reproducible than a controlled bibliographic database. Although the final search was updated on September 6, 2026, rapid publication and evolving regulatory positions will shorten the currency of some conclusions.

The evidence base is affected by publication bias, commercial confidentiality, selective disclosure of successful programs, and limited access to negative predictions. Some fields rely heavily on retrospective benchmarks; others report prospective hits without full denominators or economic comparisons. Company involvement does not invalidate a study, but it increases the importance of transparent methods, independent replication, and disclosure of conflicts. These limitations reduce confidence in universal claims about cost and time while remaining compatible with bounded conclusions about validated uses.

Finally, a narrative evidence map cannot determine a pooled effect size or comparative ranking of AI systems. Conclusions must be adapted to the disease, modality, dataset, laboratory, and regulatory context of each use. The maturity ratings are interpretive summaries, not formal grades of therapeutic efficacy (Zhang et al., 2025).

19. Conclusion

Artificial intelligence can improve drug discovery when it increases the quality of a defined decision and survives appropriate validation. Its strongest demonstrated contributions include representation learning, structure prediction, search-space compression, prospective prioritization, multi-objective hypothesis generation, and integration of heterogeneous evidence. These capabilities do not collapse the distance between a model output and a medicine. Halicin is an AI-prioritized preclinical antibacterial candidate, not an approved AI-discovered drug; abaucin provides a second preclinical prospective exemplar; AlphaFold's structural accuracy does not establish ligand affinity or medicinal impact; and rentosertib's randomized phase 2a study establishes clinical translation, not approval or definitive efficacy. The five evidence levels and the stage-gated governance framework are author-proposed analytical tools. Responsible implementation requires realistic splits and strong baselines, complete tested denominators and failed predictions, calibrated uncertainty, reproducible experimental confirmation, substantive human oversight, lifecycle change control, and context-specific regulatory credibility. The relevant endpoint is not the volume of predictions but the production of trustworthy therapeutic evidence.

The discovery of halicin remains a compelling example of this model: deep learning identified a structurally unusual candidate, and researchers established antibacterial

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

activity through laboratory and animal studies. It demonstrates prospective prioritization, not the automatic creation of a clinically proven medicine (Stokes et al., 2020).

Reliable adoption requires representative data, appropriate splits, uncertainty, strong baselines, prospective validation, provenance, interdisciplinary review, security, and life-cycle monitoring. Regulatory and scientific claims must specify the model's context of use and separate computational, preclinical, clinical, and approval milestones (FDA, 2025; NIST, 2023).

AI should therefore be understood as an accelerator of hypothesis generation and learning, not a substitute for medicinal chemistry, pharmacology, toxicology, clinical research, or regulatory judgment. The most important advance will be a discovery system that learns faster while preserving evidentiary rigor, reproducibility, and patient safety.

Ethical Approval

Ethical approval was not required because this literature review did not involve human participants, identifiable patient data, or experimental interventions. Ethical and regulatory issues discussed in the article refer to the reviewed technologies rather than procedures conducted by the author.

Funding

No external funding was received for the preparation of this article.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability

No new datasets were generated or analyzed. All evidence used in this review is contained in the cited publications and institutional documents. The source-evidence extraction matrix is provided as Supplementary Table S1.

Author Contributions

Paola Carolina Rainer da Silva: conceptualization, methodology, investigation, evidence appraisal, writing—original draft, writing—review and editing, and final approval of the manuscript.

Declaration of Generative AI Use

Generative AI was used during revision for language editing, structural consistency checking, and assistance in organizing responses to editorial requests. The author independently verified all scientific claims, source attributions, terminology, and

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

references; made the final decisions regarding content; and accepts full responsibility for the manuscript.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Table 1. Characteristics of the Analytical Evidence Map

Source(s)	Application/source type	Validation level	Principal contribution	Main limitation/maturity
Paul et al. (2010); Scannell et al. (2012)	Pharmaceutical R&D productivity	Industry-level synthesis	Attrition and productivity context	Historical, not AI-specific; contextual
Lavecchia (2015); Reker & Schneider (2015)	Screening/active learning	Methods and review evidence	Screening and iterative selection mechanisms	Predates foundation-model era; moderate
Wu et al. (2018); Yang et al. (2019)	Molecular benchmarks/representations	Benchmark evaluation	Dataset-aware molecular prediction	Benchmark-to-practice gap; moderate
Stokes et al. (2020) and published correction	Halicin; AI-prioritized antibacterial	Prospective in vitro and animal	Unusual scaffold prioritized and experimentally validated	Preclinical; stronger for prioritization, not therapy
Liu et al. (2023)	Abaucin; AI-prioritized antibacterial	Prospective laboratory	Pathogen-selective candidate	Preclinical/narrow context; stronger for prioritization
Jumper et al. (2021); Varadi et al. (2022); Akdel et al. (2022); Abramson et al. (2024)	Protein/complex structure	Benchmarks, resource, community assessment	High-accuracy structural hypotheses and expanded access	Affinity, dynamics, druggability limits
Elton et al. (2019); Schneider et al. (2020); Walters & Murcko (2020)	AI-designed/generated molecules	Critical synthesis	Methods and medicinal-chemistry evaluation	Limited prospective comparison; emerging-to-moderate
MacLeod et al. (2020)	Closed-loop laboratory	Prospective autonomous experimentation	Iterative model-experiment automation	Materials-domain exemplar; emerging for drugs

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Source(s)	Application/source type	Validation level	Principal contribution	Main limitation/maturity
Swanson et al. (2024)	ADMET-AI	Multi-dataset benchmark	Open scalable multi-endpoint ADMET prioritization	Retrospective benchmark; endpoint-dependent
Pushpakom et al. (2019); Zitnik et al. (2018)	Repurposing/graph learning	Review and retrospective	Integrates network and pharmacological hypotheses	Confounding and task specificity; emerging-to-moderate
Xu et al. (2025)	Rentosertib; AI-assisted program	Randomized phase 2a	Clinical translation with 71 participants	Exploratory; not confirmatory or regulatory success
FDA (2025); FDA & EMA (2026); NIST (2023); USPTO (2025)	Governance/regulatory/legal	Official frameworks/guidance	Context of use, credibility, lifecycle, inventorship	Normative/contextual, not therapeutic proof

Note. Core analytical and contextual sources are shown by decision relevance. The Stokes et al. correction is explicitly incorporated.

Table 2. Evidence Maturity by Pipeline Domain

Domain	Narrative maturity	Most persuasive evidence	Unresolved question
Protein/complex structure	Stronger for structure; moderate for drug decisions	Independent benchmarks and structural resources	Conformation, dynamics, affinity, prospective medicinal impact
Virtual screening/prioritization	Moderate	Prospective hits including halicin and abaucin	Generalization, tested denominators, comparative hit rate
Generative molecular design	Emerging-to-moderate	Synthesized candidates and phase 2a clinical exemplar	Replication, comparative success, total cost, long-term safety
ADMET/toxicity	Moderate, endpoint-dependent	Multi-dataset benchmarks and selected experimental validation	Rare toxicities, protocol shift, species translation, calibration

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Domain	Narrative maturity	Most persuasive evidence	Unresolved question
Drug repurposing	Emerging-to-moderate	Convergent network, pharmacological, and trial evidence	Confounding, historical bias, exposure, clinical benefit
Closed-loop laboratories	Emerging for pharmaceuticals	Prospective autonomous experimentation mainly outside drugs	Pharmaceutical quality systems, transfer, scalability
Clinical-development support	Emerging and use-specific	Randomized trial, biomarker, and operational evidence	External validity, endpoint integrity, equity, regulatory acceptance

Note. Maturity labels apply the criteria in Section 2.4. They are narrative confidence judgments, not GRADE or regulatory grades. Structural-prediction accuracy is not equivalent to validated affinity or medicinal impact.

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Table 3. Stage-Gated Governance for AI-Augmented Drug Discovery

Stage/context	Responsible actors	Minimum evidence/control	Monitoring/accountability
Exploratory research	Data steward; biologist; data scientist	Provenance, license/consent, relevant split, baseline, applicability domain	Dataset/model versioning; technical review
Development decision	Medicinal chemist; pharmacologist; toxicologist; project lead	Uncertainty, feasibility, orthogonal assays, replication, ADMET, complete tested denominator	Override rationale; stage-gate minutes; revalidation after change
Clinical decision	Clinician; statistician; sponsor; ethics team	Validated endpoint, population relevance, bias analysis, prespecified threshold	Trial oversight; drift and incident review
Regulatory submission	Sponsor; quality unit; regulatory specialists	Defined context of use, risk-based credibility, traceability, acceptance criteria	Change control and regulator-ready documentation

Note. Author synthesis based on the reviewed literature; governance requirements increase with context-specific risk.

Table 4. Evidence Requirements Across the AI-Enabled Drug Discovery Pipeline

Stage	AI contribution	Required evidence
Target selection	AI-enabled integration of genetics, omics, networks, literature	Causal rationale, independent datasets, perturbational and disease-relevant validation
Hit discovery	AI-prioritized screening, docking rescoring, active learning	Prospective hit yield, strong baselines, orthogonal assays, counter-screens, novel scaffolds, complete tested set

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Stage	AI contribution	Required evidence
Lead optimization	AI-designed/generated alternatives and multi-property prediction	Synthesizability, experimental potency/selectivity, endpoint-specific ADMET, uncertainty, strong baselines
Translation	AI-supported dose, biomarker, trial, or regulatory analysis	Predefined context of use, clinical validation, traceability, monitoring, regulatory credibility

Note. Author synthesis based on the reviewed literature; governance requirements increase with context-specific risk.

Supplementary Table S1. Source-Evidence Extraction Matrix

Source cluster	Role	Decision/domain	Evidence level	Principal limitation
Stokes et al. (2020) + correction	Primary	AI-prioritized antibacterial	Prospective in vitro/in vivo	Preclinical; correction required for provenance
Liu et al. (2023)	Primary	AI-prioritized antibacterial	Prospective laboratory	Narrow pathogen context
Jumper (2021); Varadi (2022); Akdel (2022); Abramson (2024)	Primary/resource/assessment	Structural hypotheses	Benchmark/community	Not affinity or therapeutic validation
MacLeod et al. (2020)	Primary transfer exemplar	Closed-loop automation	Prospective materials	Not pharmaceutical end-to-end evidence
Swanson et al. (2024)	Primary benchmark/tool	ADMET prioritization	41-dataset/leaderboard	Retrospective and endpoint-dependent
Xu et al. (2025)	Primary clinical	AI-assisted program	Randomized phase 2a; n=71	Exploratory; no approval or definitive efficacy
FDA (2025)	Regulatory	Formal regulatory decision	Draft Level 1 guidance	Nonbinding; excludes discovery-stage models

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Source cluster	Role	Decision/domain	Evidence level	Principal limitation
FDA & EMA (2026)	Regulatory principles	Lifecycle governance	Joint guiding principles	Not binding approval criteria
NIST (2023); WHO (2021)	Governance/ethics	Risk, security, equity	Framework/guidance	Normative rather than therapeutic evidence
USPTO (2025)	Legal authority	Inventorship	Revised U.S. guidance	U.S.-specific; not globally uniform
Henderson et al. (2020)	Methodological	Energy/carbon reporting	Direct reporting framework	General ML rather than drug-specific
Critical reviews and benchmarks	Analytical/contextual	Pipeline-wide	Secondary/retrospective	Heterogeneity and benchmark-to-practice gap

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C. C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 630(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Akdel, M., Pires, D. E. V., Pardo, E. P., Jänes, J., Zalevsky, A. O., Mészáros, B., Bryant, P., Good, L. L., Lasker, K., Pozzati, G., Shenoy, A., Zhu, W., Kundrotas, P., Serra, V. R., Rodrigues, C. H. M., Dunham, A. S., Burke, D., Borkakoti, N., Velankar, S., … Beltrao, P. (2022). A structural biology community assessment of AlphaFold2 applications. Nature Structural & Molecular Biology, 29, 1056–1067. https://doi.org/10.1038/s41594-022-00849-w

Bender, A., & Cortés-Ciriano, I. (2021a). 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 Discovery Today, 26(2), 511–524. https://doi.org/10.1016/j.drudis.2020.12.009

Bender, A., & Cortés-Ciriano, I. (2021b). Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discovery Today, 26(4), 1040–1052. https://doi.org/10.1016/j.drudis.2020.11.037

Elton, D. C., Boukouvalas, Z., Fuge, M. D., & Chung, P. W. (2019). Deep learning for molecular design—A review of the state of the art. Molecular Systems Design & Engineering, 4(4), 828–849. https://doi.org/10.1039/C9ME00039A

Food and Drug Administration. (2025, January). Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: Draft guidance for industry and other interested parties (Draft Level 1; not for implementation). U.S. Department of Health and Human Services. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

Food and Drug Administration, & European Medicines Agency. (2026). Guiding principles of good AI practice in drug development. https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development

Fu, C., & Chen, Q. (2025). The future of pharmaceuticals: Artificial intelligence in drug discovery and development. Journal of Pharmaceutical Analysis, 15, 101248. https://doi.org/10.1016/j.jpha.2025.101248

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Huang, K., Xiao, C., Glass, L. M., & Sun, J. (2022). Artificial intelligence for therapeutic science. Nature Chemical Biology, 18, 1033–1036. https://doi.org/10.1038/s41589-022-01131-2

Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: Methods and applications. Drug Discovery Today, 20(3), 318–331. https://doi.org/10.1016/j.drudis.2014.10.012

Liu, G., Catacutan, D. B., Rathod, K., Swanson, K., Jin, W., Mohammed, J. C., Chiappino-Pepe, A., Syed, S. A., Fragis, M., Rachwalski, K., Magolan, J., Surette, M. G., Coombes, B. K., Jaakkola, T., Barzilay, R., Collins, J. J., & Stokes, J. M. (2023). Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. Nature Chemical Biology, 19(11), 1342–1350. https://doi.org/10.1038/s41589-023-01349-8

MacLeod, B. P., et al. (2020). Self-driving laboratory for accelerated discovery of thin-film materials. Science Advances, 6(20), eaaz8867. https://doi.org/10.1126/sciadv.aaz8867

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Niazi, S. K., & Mariam, Z. (2025). Artificial intelligence in drug development: Reshaping the therapeutic landscape. Therapeutic Advances in Drug Safety, 16, 20420986251321704. https://doi.org/10.1177/20420986251321704

Paul, S. M., et al. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery, 9, 203–214. https://doi.org/10.1038/nrd3078

Pushpakom, S., et al. (2019). Drug repurposing: Progress, challenges and recommendations. Nature Reviews Drug Discovery, 18, 41–58. https://doi.org/10.1038/nrd.2018.168

Reker, D., & Schneider, G. (2015). Active-learning strategies in computer-assisted drug discovery. Drug Discovery Today, 20(4), 458–465. https://doi.org/10.1016/j.drudis.2014.12.004

Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. Nature Reviews Drug Discovery, 11, 191–200. https://doi.org/10.1038/nrd3681

Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., Jr., Fisher, J., Jansen, J. M., Duca, J. S., Rush, T. S., Zent, E., & Hill, J. E. (2020). Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery, 19, 353–364. https://doi.org/10.1038/s41573-019-0050-3

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stokes, J. M., et al. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021 The publisher's correction linked to the article record was consulted in the present review.

Varadi, M., et al. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research, 50(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Vamathevan, J., et al. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18, 463–477. https://doi.org/10.1038/s41573-019-0024-5

Walters, W. P., & Murcko, M. (2020). Assessing the impact of generative AI on medicinal chemistry. Nature Biotechnology, 38, 143–145. https://doi.org/10.1038/s41587-020-0418-2

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. Journal of Advanced Nursing, 52(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. https://www.who.int/publications/i/item/9789240029200

Xu, Z., Ren, F., Wang, P., Cao, J., Tan, C., Ma, D., Zhao, L., Dai, J., Ding, Y., Fang, H., Li, H., Liu, H., Luo, F., Meng, Y., Pan, P., Xiang, P., Xiao, Z., Rao, S., Satler, C., … Zhavoronkov, A. (2025). A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: A randomized phase 2a trial. Nature Medicine, 31(8), 2602–2610. https://doi.org/10.1038/s41591-025-03743-2

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. Chemical Science, 9(2), 513–530. https://doi.org/10.1039/C7SC02664A

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237

Zhang, K., Yang, X., & Yang, S. (2025). Artificial intelligence in drug development. Nature Medicine, 31, 45–59. https://doi.org/10.1038/s41591-024-03434-4

CC BY

Nexus Science Review
INTERNATIONAL JOURNAL OF INTERDISCIPLINARY RESEARCH

Received: 14/08/2026 - **Accepted:** 13/09/2026
Vol: 01.05 - October
DOI: 10.66104/337xw232
Pages: 1-32

Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics, 34(13), i457–i466. https://doi.org/10.1093/bioinformatics/bty294

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248), 1–43. https://www.jmlr.org/papers/v21/20-312.html

Swanson, K., Walther, P., Leitz, J., Mukherjee, S., Wu, J. C., Shivnaraine, R. V., & Zou, J. (2024). ADMET-AI: A machine learning ADMET platform for evaluation of large-scale chemical libraries. Bioinformatics, 40(7), btae416. https://doi.org/10.1093/bioinformatics/btae416

United States Patent and Trademark Office. (2025, November 26). Revised inventorship guidance for AI-assisted inventions. https://www.uspto.gov/subscription-center/2025/revised-inventorship-guidance-ai-assisted-inventions

CC BY