Trustworthy AI for Computational Drug Discovery: A Realist Review of Explainability, Reproducibility, Validation, and Regulatory Confidence

Sara Al-Fahd[1]*, Noura Al-Khalifa[1], Omar Al-Turki[2]

[1]Department of AI-Enabled Drug Repurposing from Desert Plants, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia.

[2]Department of Computational Pharmacology and Docking Studies, College of Pharmacy, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

ABSTRACT

Artificial intelligence is increasingly embedded in computational drug discovery, yet predictive capability alone does not establish whether an AI system is scientifically credible, reproducible, generalizable, or sufficiently governed to support consequential decisions. This realist review examines when, how, why, and under what conditions AI becomes trustworthy enough to contribute to computational drug-discovery workflows, with particular attention to explainability, reproducibility, validation, and regulatory confidence. Rather than treating these domains as independent technical attributes, the review develops context–mechanism–outcome configurations to explain how trust-building processes are activated, inhibited, or redirected by data provenance, chemical-space coverage, assay quality, workflow transparency, stakeholder expertise, validation design, decision consequence, and governance maturity. The synthesis shows that explainability may strengthen scrutiny, debugging, plausibility assessment, and human challenge when explanations are sufficiently faithful, stable, domain-relevant, and competently interpreted; under weaker conditions, the same explanatory layer may generate plausibility bias, automation bias, and false reassurance. Reproducibility can enable independent verification and error discovery when computational environments, preprocessing, versioning, randomness, and workflow provenance are reconstructable, but reproducible execution cannot transform invalid assumptions or leakage-contaminated evidence into valid science. Validation confidence depends on the distinctness and relevance of the challenge context, while uncertainty estimates require task-specific evaluation and decision-sensitive interpretation. Regulatory confidence is similarly conditional on traceability, intended-use clarity, evidence provenance, change control, accountable oversight, and risk-proportionate validation rather than transparency alone. The refined programme theory therefore conceptualizes trustworthy AI as a bounded and revisable outcome emerging from mutually checking mechanisms operating within supportive contexts. For computational drug discovery, implementation should prioritize independent challenge, transparent provenance, calibrated uncertainty awareness, and governance intensity proportionate to decision risk.

Key Words: *Trustworthy artificial intelligence, Computational drug discovery, Explainable AI, Reproducibility, External validation, Regulatory confidence*

eIJPPR 2025; 14(2):23-39

HOW TO CITE THIS ARTICLE: Al-Fahd S, Al-Khalifa N, Al-Turki O. Trustworthy AI for Computational Drug Discovery: A Realist Review of Explainability, Reproducibility, Validation, and Regulatory Confidence. Int J Pharm Phytopharmacol Res. 2025;15(2):23-39. https://doi.org/10.51847/awiGUCa4tU

INTRODUCTION

Artificial intelligence now contributes to a broad range of computational drug-discovery activities, including molecular property prediction, target-related inference, virtual screening, ADMET estimation, toxicity prediction, molecular generation, and biomolecular modeling.

Corresponding author: Sara Al-Fahd
Address: Department of AI-Enabled Drug Repurposing from Desert Plants, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia.
E-mail: ✉ s.alfahd@ksu.edu.sa
Received: 04 December 2024; **Revised:** 02 April 2025; **Accepted:** 05 April 2025

Machine-learning methods have consequently become relevant at multiple stages of discovery and development, although the evidentiary meaning of a prediction varies substantially with task, data source, and decision context [1]. Reviews of the field similarly show an expanding methodological landscape in which deep learning and related AI approaches are used to augment screening, prioritization, design, and development processes rather than perform one uniform function [2]. This heterogeneity is central to trustworthiness because a model supporting exploratory hypothesis generation is exposed to different consequences, evidentiary requirements, and tolerable failure modes than a system influencing expensive experimental progression or high-consequence safety decisions.

The growth of predictive capability has intensified a parallel concern: computational performance can improve without a corresponding increase in scientific credibility. AI-era drug design is increasingly characterized by interaction among computational models, medicinal chemistry judgment, biological experimentation, and iterative decision processes, making the quality of integration at least as important as the presence of an advanced architecture [3]. A system may rank molecules accurately within a familiar benchmark yet remain poorly understood outside that benchmark; another may generate apparently plausible structures while offering little evidence that its learned relationships are transferable to future chemical or biological contexts. Trust therefore cannot be reduced to whether a model is sophisticated, whether its outputs appear chemically intuitive, or whether it outperforms a comparator on a retrospective test set.

This tension is reinforced by critical analyses of what counts as realistic AI impact in drug discovery. Strong computational results may coexist with weak linkage to genuine discovery decisions, limited experimental challenge, inappropriate comparators, or uncertainty about whether measured improvement translates into practical value [4]. High predictive performance can therefore be a necessary contribution for some tasks without being sufficient evidence of trustworthiness. Scientific confidence additionally depends on what information entered the model, whether the evaluation design reflects the intended use, whether failure boundaries are visible, whether another group can reconstruct the workflow, and whether the evidentiary claim is proportionate to the strength of validation.

Chemical and biological data create distinctive reasons for this caution. Assay heterogeneity, biased measurement processes, sparse labels, chemical redundancy, unbalanced endpoints, uncertain target relationships, and limited coverage of the practically relevant chemical space can condition both apparent performance and downstream transportability [5]. The same modeling mechanism may consequently support credible prioritization in one setting and false confidence in another. A realist review is appropriate because it asks not merely whether explainability, reproducibility, validation, or governance "works," but what works, for whom, under which drug-discovery conditions, through which mechanisms, and with what outcomes. The aim of this review is therefore to develop and refine an evidence-grounded programme theory explaining how trustworthy AI outcomes emerge—or fail to emerge—through interactions among technical context, scientific workflow, validation design, human response, and regulatory or organizational conditions.

Trustworthy AI as a drug discovery challenge

For computational drug discovery, trustworthy AI should be defined as a context-dependent capacity to support decisions with evidence that is sufficiently credible, inspectable, reproducible, uncertainty-aware, and appropriately validated for the intended use. This definition is narrower and more operational than a generic ethical characterization of "responsible" AI. It includes technical performance but does not equate performance with trust; it includes explainability but does not presume that an explanation establishes model correctness; and it includes governance while recognizing that procedural controls cannot substitute for scientific validity. Drug-discovery-specific work on explainable AI illustrates the potential value of explanations for inspecting model behavior, communicating patterns, and linking predictions to chemically meaningful features, while also emphasizing substantial methodological and interpretive limitations [6]. Trustworthiness is therefore best treated as an emergent property of an evidence system rather than an intrinsic label attached to a model.

The challenge becomes especially acute when explanation plausibility is mistaken for evidentiary strength. A heat map, attributed substructure, attention pattern, or counterfactual example may appear intelligible to a scientist without being faithful to the computation that produced the prediction. Critical work in high-stakes AI has warned that current explanation practices can create unwarranted confidence when users infer reliability from persuasive narratives rather than independent evidence of validity [7]. In drug discovery, this problem may be amplified when an explanation resembles accepted medicinal-chemistry intuition: agreement with prior expectations can facilitate scrutiny, but it can also trigger confirmation bias. The relevant trust mechanism is not explanation visibility itself; it is competent, critical interrogation of an explanation whose fidelity, stability, domain relevance, and limitations have been considered.

ISSN (Online) 2249-6084 (Print) 2250-1029

Reproducibility introduces a different but equally conditional dimension. A computational claim becomes more open to independent scrutiny when investigators can reconstruct the data processing, software environment, model configuration, randomness controls, evaluation protocol, and analytical choices that generated it. Reproducibility standards proposed for machine learning in the life sciences emphasize that code availability alone is insufficient when data access, dependencies, preprocessing, hyperparameters, or execution environments remain opaque [8]. For computational drug discovery, this distinction is consequential because seemingly minor differences in molecular standardization, train–test construction, feature generation, or tuning can alter results. Reproducibility can therefore activate independent verification and error detection, yet it does not independently prove that the task formulation or validation design was scientifically appropriate.

Validation is similarly vulnerable to false confidence when evaluation conditions fail to represent the information structure of intended use. Leakage can occur through preprocessing, target construction, duplicated or related entities, inappropriate temporal ordering, or other pathways that expose the model to information unavailable in the future decision setting. Methodological analysis of leakage in machine-learning-based science shows how such pathways can inflate estimated performance while weakening reproducibility and scientific inference [9]. In molecular prediction, the problem intersects with chemical redundancy, scaffold relationships, repeated benchmark use, and distribution shift. Random splitting is not inherently invalid, but it may be an insufficient challenge when near-neighbor compounds or related information structures occur across partitions and the intended claim concerns novel chemical space.

Stakeholder diversity further conditions what "trustworthy" means in practice. Computational scientists may prioritize reproducible evaluation and diagnostic understanding; medicinal chemists may require chemically actionable reasoning; biologists and toxicologists may focus on assay relevance, endpoint validity, and uncertainty; pharmaceutical decision-makers may need evidence proportional to development cost and consequence; quality teams may emphasize traceability and change control; and regulators may require intended-use clarity, evidence provenance, and inspectable risk management. These perspectives are not interchangeable, and a model may be trusted for one bounded purpose while remaining unsuitable for another. **Table 1** maps the major trustworthiness challenges in computational drug discovery to their enabling or constraining contexts, underlying mechanisms, potential outcomes, and evidentiary requirements.

Table 1. Trustworthy AI Challenges in Computational Drug Discovery: Context Conditions, Trust-Building and Trust-Undermining Mechanisms, Potential Outcomes, Validation Requirements, and Risks of False Confidence

Trustworthiness challenge	Drug-discovery context	Enabling context	Constraining context	Mechanism	Potential positive outcome	Potential negative outcome	Validation requirement	Risk of false confidence
Predictive performance without scientific credibility	Molecular property prediction, target prediction, virtual screening	Clear intended use; relevant endpoints; representative data; appropriate comparators	Weak endpoint validity; biased labels; unclear decision purpose	Independent scientific challenge of performance claims	More credible model selection and prioritization	Performance interpreted as proof of usefulness	Task-aligned internal and independent validation	High when benchmark improvement is detached from intended use
Explanation plausibility	Compound activity prediction, GNN interpretation, molecular prioritization	Faithful explanation method; domain expertise; stability checks	Post hoc rationalization; weak fidelity; confirmation bias	Explanation-enabled scrutiny and plausibility checking	Error detection; model debugging; better human challenge	Explanation overtrust; illusory transparency	Explanation fidelity, stability, robustness, and decision-utility assessment	High when intuitive explanations are treated as evidence of correctness
Chemical applicability limits	Extrapolation to new scaffolds, rare chemotypes, shifted series	Adequate chemical-space coverage; explicit applicability domain	Sparse coverage; activity cliffs; hidden distribution shift	Boundary awareness and cautious reliance	More credible prioritization within supported space	Overconfident extrapolation	Scaffold-aware, shifted-context, temporal, or external challenge	High when in-domain accuracy is generalized beyond support

							where appropriate	
Data integrity and provenance	QSAR, ADMET, toxicity, multimodal prediction	Traceable sources; assay metadata; consistent curation	Assay heterogeneity; uncertain labels; undocumented merging	Data-lineage scrutiny and error tracing	Better interpretation of model limitations	Hidden bias and untraceable error	Provenance audit and sensitivity analysis	High when dataset size is mistaken for representativeness
Reproducibility	Model development and comparative benchmarking	Code, data, preprocessing, dependencies, versions, seeds, hyperparameters documented	Proprietary opacity; missing preprocessing; environment mismatch	Independent rerun and workflow reconstruction	Verification and error discovery	Replication failure or unverifiable claims	Independent reconstruction or controlled reproducibility assessment	High when code availability alone is treated as reproducibility
Leakage and benchmark contamination	Retrospective molecular ML evaluation	Leakage-aware splitting; partition audit; temporal and entity controls	Related compounds across splits; preprocessing contamination; repeated tuning	Independent challenge of information pathways	More realistic performance estimates	Inflated benchmark confidence	Split rationale, leakage audit, untouched challenge data	Very high because apparent performance may remain numerically impressive
Uncertainty estimation	Molecular property prediction and prioritization	Task-specific calibration; explicit decision rule; shift awareness	Miscalibration; unsupported extrapolation; poorly interpreted scores	Uncertainty awareness changes reliance or escalation	More cautious prioritization and better triage	Numerical confidence mistaken for reliability	Calibration, stress testing, and evaluation under relevant shift	High when uncertainty is displayed but not validated
External generalization	Cross-dataset, temporal, institutional, or chemical-space transfer	Genuinely distinct and decision-relevant external context	Nominally external data sharing the same biases or near-duplicates	External challenge	Bounded generalization confidence	False claims of universal transportability	Explicit assessment of contextual distinctness	High when "external" is treated as a binary label
Prospective or experimental confirmation	ADME prediction, candidate prioritization, generative design	Future-facing cases; pre-specified workflow; empirical testing	Selective follow-up; changed protocols; weak comparator	Exposure to future or physical evidence	Stronger translational confidence	Overinterpretation of isolated successes	Prospective and/or experimental confirmation appropriate to claim	Moderate to high when confirmation is narrow but claims are broad
Human oversight	Expert review of model-supported decisions	Domain expertise; authority to challenge; escalation pathways	Automation bias; workload; ambiguous accountability	Human error interception	Better scrutiny and cautious escalation	Rubber-stamping and overreliance	Evaluation of oversight role and decision process	High when a nominal "human in the loop" lacks practical authority
Traceability and governance	Pharmaceutical implementation and regulated development	Intended-use clarity; ownership; auditability; change control	Fragmented accountability; undocumented updates; governance theater	Evidence reconstruction and accountable control	Implementation readiness and stronger audit confidence	Uncontrolled drift and regulatory uncertainty	Lifecycle documentation and reassessment after material change	High when documentation volume is mistaken for scientific validity

Realist review logic

Realist review is suited to trustworthy AI because the central problem is explanatory rather than merely classificatory. The objective is not to count how often explainability, reproducibility, uncertainty quantification, or external validation appears in the literature, but to

determine how particular conditions enable these interventions to trigger responses that alter confidence. A context–mechanism–outcome configuration is therefore used as a provisional causal explanation: a context shapes whether a mechanism is activated, and that mechanism contributes to an outcome under specified conditions. Methodological work on realist configurations emphasizes that configuration design must serve an explanatory purpose rather than function as a decorative arrangement of variables [10]. In this review, CMO reasoning is consequently used to compare positive pathways, mechanism failures, negative cases, and boundary conditions.

Context is defined as the technical, scientific, data, workflow, organizational, human, governance, and regulatory conditions within which an AI system and its users operate. It includes such factors as data provenance, assay heterogeneity, chemical-space coverage, stakeholder expertise, workflow maturity, decision consequence, institutional infrastructure, validation design, and intended use. Realist-method scholarship cautions against treating context as a static list of background variables because contexts can be multilevel, relational, and causally consequential for whether mechanisms activate [11]. Thus, "code availability" may be enabling in an environment with accessible data and adequate computational expertise but insufficient in a setting where preprocessing remains undocumented or dependencies cannot be reconstructed.

A mechanism is treated as the process or response activated within a context rather than simply the technical component that has been introduced. An explanation method is therefore not itself the complete mechanism; a relevant mechanism may be explanation-enabled scrutiny, confirmation bias, or overtrust. Similarly, an external dataset is not a mechanism; the mechanism is external challenge when the dataset exposes a model to meaningfully distinct conditions and thereby tests assumptions about transportability. Theory-oriented analysis of AI uptake in pathology illustrates the value of examining how organizational, workflow, user, and technological conditions interact to shape implementation responses and outcomes [12]. Although evidence transferred from adjacent biomedical AI must be qualified, this logic is directly useful for studying why apparently similar trust interventions may succeed in one drug-discovery setting and fail in another.

The review process therefore uses iterative and purposive searching to test an initial programme theory rather than claim exhaustive retrieval. Theory-seeking searches identify candidate explanations; mechanism-seeking searches examine how scrutiny, replication, uncertainty awareness, external challenge, traceability, and human oversight are expected to operate; contradictory-evidence searches identify leakage, explanation overtrust, benchmark inflation, distribution shift, and implementation failure. Relevance is assessed by asking whether a source can inform, support, refine, challenge, or reject a CMO configuration, whereas rigor is assessed by asking whether the evidence is sufficiently credible for the specific inference being drawn. The same intervention can consequently generate different outcomes: explanation may support model debugging in a chemically interpretable task but produce false reassurance when fidelity is weak; reproducibility may enable independent error detection but merely reproduce a leakage-contaminated analysis; and governance may strengthen accountability when operationalized while degenerating into documentation theater when scientific validity remains unchallenged.

Explainability and reproducibility mechanisms

Explainability mechanisms occupy a prominent but contested position in trustworthy AI. Intrinsic interpretability, post hoc feature attribution, chemical substructure attribution, graph explanations, counterfactual explanations, example-based explanations, concept-level methods, and attention-based interpretations differ in what they reveal and what claims they can support. A crucial distinction is whether a model is directly interpretable in a decision-relevant sense or whether an auxiliary procedure is attempting to explain a complex black box after prediction. Arguments for interpretable modeling in high-stakes decisions warn that post hoc explanation can obscure this distinction and may provide an appearance of understanding without equivalent access to model logic [13]. In computational drug discovery, the appropriate choice depends on task complexity, available alternatives, scientific purpose, and the cost of misleading interpretation.

Feature and substructure attributions can nevertheless activate useful mechanisms when they expose model sensitivities to scientifically informed scrutiny. In neural models for chemistry, attribution has been used to interrogate features associated with predicted binding behavior and to generate hypotheses about what the model has learned [14]. The appropriate outcome is not causal proof but increased inspectability: a chemist may identify an implausible dependency, compare influential regions with known structure–activity reasoning, or discover that a model relies on a shortcut. Counterfactual explanations provide a complementary mechanism by asking what chemically meaningful change would alter a prediction, thereby making local prediction sensitivity more contrastive and potentially actionable [15]. Their utility, however, depends on whether proposed changes are feasible, sufficiently close to the original molecule, stable

across model perturbations, and relevant to the decision being made.

The critical realist distinction is between trust-building mechanisms and trust-signaling artifacts. An explanation becomes trust-building when it enables competent challenge, reveals errors, clarifies boundaries, or supports a decision that remains open to independent verification. It becomes trust-signaling when a visually persuasive output merely increases subjective confidence. Quantitative evaluation of molecular graph explanations demonstrates why explanation quality should be challenged rather than presumed, because explainer performance varies with model, task, and evaluation criterion [16]. Domain-aware approaches that incorporate chemically meaningful substructure information may improve the alignment of explanatory outputs with relevant molecular patterns, but their value remains conditional on the appropriateness of the prior knowledge and the validity of the evaluation context [17]. Thus, a plausible explanation may be useful, yet plausibility is not equivalent to fidelity, and neither property independently validates the underlying prediction.

Figure 1 illustrates how explainability and reproducibility mechanisms can either strengthen scientific confidence or generate false reassurance depending on data quality, workflow transparency, explanation fidelity, stakeholder expertise, and verification conditions.

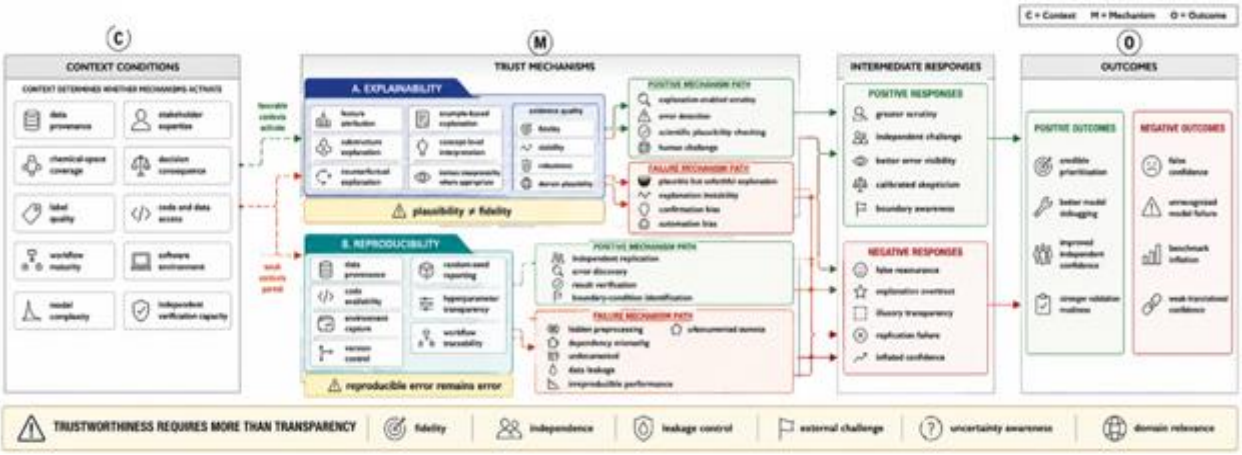

Figure 1. Conditional Trust Pathways for Explainability and Reproducibility in AI-Enabled Drug Discovery

Alt-text description

A wide left-to-right conditional pathway begins with contextual conditions such as data provenance, chemical-space coverage, workflow transparency, explanation fidelity, code availability, and stakeholder expertise. These conditions activate or inhibit explainability and reproducibility mechanisms. Positive branches lead to model scrutiny, error detection, independent replication, and calibrated confidence, whereas failure branches lead to plausible but unfaithful explanations, automation bias, replication failure, hidden leakage, and false confidence.

Reproducibility mechanisms address a related but distinct problem: whether another investigator can reconstruct and challenge the computational pathway that produced a result. Benchmark standardization can support comparison by making tasks, datasets, and evaluation procedures more explicit, but the meaning of benchmark evidence remains dependent on split strategy, endpoint, dataset structure, and the generalization claim being made [18]. Robust reproducibility therefore extends beyond making a script public. It requires transparent molecular standardization, preprocessing history, feature generation, data partitions, hyperparameters, random seeds, dependency versions, computational environments, stopping rules, and benchmark documentation. Under supportive conditions, these resources activate independent rerun, discrepancy detection, and boundary-condition discovery.

The importance of evaluation context becomes particularly visible when benchmark structure rewards memorization rather than transportable learning. Empirical analysis of ligand-based classification has shown that some benchmarks can favor recognition of related compounds instead of generalization to genuinely distinct chemical matter [19]. This failure pathway is not simply a “bad model” problem; it is a context–mechanism interaction. When chemically related observations cross evaluation boundaries, a model may exploit similarity, the benchmark may reward that behavior, and investigators may infer generalization from an insufficient challenge. Comparable concerns arise from hidden preprocessing, repeated benchmark tuning, undocumented selection decisions, and data leakage. Reproducibility is valuable here because it can expose these pathways, but it cannot guarantee that

ISSN (Online) 2249-6084 (Print) 2250-1029

they will be recognized unless the evaluation logic itself is critically inspected.

Explainability and reproducibility should consequently be understood as mutually checking but non-substitutable mechanisms. Explanation can expose suspicious model dependencies while reproducibility enables others to reconstruct the conditions under which those dependencies appeared. Conversely, a reproducible workflow may consistently generate a misleading explanation, and a faithful explanation may describe a model trained on biased or leakage-contaminated data. The strongest trust pathway therefore combines explanation quality assessment with transparent provenance, independent rerun, explicit generalization assumptions, leakage control, and scientific challenge. The corresponding failure pathway combines persuasive explanation, opaque preprocessing, unverifiable tuning, and uncritical stakeholder reliance. **Table 2** compares major explainability and reproducibility mechanisms according to their intended trust function, contextual requirements, failure modes, and evidentiary limits.

Table 2. Explainability and Reproducibility Mechanisms for Trustworthy AI in Computational Drug Discovery: Intended Functions, Contextual Preconditions, Scientific Benefits, Failure Modes, and Validation Safeguards

Mechanism	Primary trust function	Typical computational context	Required enabling condition	Potential scientific benefit	Major failure mode	Risk of overinterpretation	Verification requirement	Implementation safeguard
Intrinsic interpretability	Make model structure or decision logic directly inspectable	Simpler predictive models; constrained high-consequence tasks	Interpretable model must retain adequate task relevance	Directer scrutiny of relationships and decision structure	Simplified model may omit important complexity	Assuming interpretability guarantees validity	Independent predictive and scientific validation	Document why interpretable form is appropriate for intended use
Feature attribution	Identify input features associated with a prediction	Molecular property prediction; neural models; descriptor-based models	Method appropriate to architecture and representation	Model debugging and sensitivity inspection	Attribution instability or method dependence	Treating importance as causality	Fidelity and perturbation-based checks	Report method assumptions and sensitivity to explainer choice
Chemical substructure attribution	Highlight molecular fragments associated with output	GNNs; QSAR; activity prediction	Chemically meaningful atom/bond representation and expert review	Detect implausible dependencies; support structure-focused scrutiny	Visually plausible but unfaithful highlighting	Equating highlighted fragment with causal mechanism	Quantitative and chemistry-aware evaluation	Separate explanatory hypothesis from validated mechanism
Graph explanation	Identify influential nodes, edges, or subgraphs	Molecular graph neural networks	Suitable explainer; explicit quality metric	Localize model-relevant structural regions	Explainer disagreement and unstable subgraphs	Assuming one graph explanation is definitive	Cross-method, fidelity, robustness, and stability testing	Preserve explainer version and evaluation protocol
Counterfactual explanation	Show changes that would alter a prediction	Molecular property and compound-activity prediction	Chemically feasible and locally relevant perturbations	Contrastive understanding and actionable inspection	Implausible or distant counterfactuals	Treating a generated change as validated optimization advice	Feasibility, proximity, stability, and prediction checks	Label counterfactuals as hypotheses requiring scientific review
Example-based explanation	Relate prediction to influential or similar examples	Similarity-rich molecular prediction	Representative reference examples and transparent similarity logic	Expose analogical basis of prediction	Example set reproduces dataset bias	Assuming similarity proves transportability	Audit example selection and chemical-space coverage	Display applicability limitations and counterexamples

ISSN (Online) 2249-6084 (Print) 2250-1029

Concept-based interpretation	Test model sensitivity to higher-level scientific concepts	Representation learning; multimodal models	Defensible concept definition and measurable concept mapping	Connect model behavior to domain constructs	Poorly specified concepts create circular interpretation	Treating concept alignment as causal understanding	Concept validity and independent robustness testing	Document concept construction and boundary conditions
Attention-based interpretation	Inspect learned weighting patterns	Transformer and attention-based molecular or biomolecular models	Evidence that attention is relevant to explanatory purpose	Hypothesis generation about information use	Attention weights are not necessarily faithful explanations	Treating attention as direct reasoning trace	Compare with alternative explanation and perturbation methods	Avoid labeling attention visualization as proof of reasoning
Explanation fidelity assessment	Determine whether explanation tracks model behavior	Any post hoc XAI workflow	Explicit fidelity criterion and relevant perturbation design	Distinguish informative explanations from persuasive artifacts	Inadequate proxy metric	Assuming one fidelity metric settles explanation validity	Multiple complementary checks where feasible	Predefine explanation claim and matching metric
Explanation stability assessment	Test whether similar cases or small perturbations yield consistent explanations	Local molecular explanation workflows	Scientifically meaningful perturbation regime	Reveal brittle explanation behavior	Stable output may still be systematically wrong	Equating stability with fidelity	Repeated and perturbation-based analysis	Report instability as a decision limitation
Data provenance capture	Enable reconstruction of training and evaluation inputs	QSAR, ADMET, toxicity, multimodal workflows	Source identifiers, transformations, inclusion logic, lineage	Error tracing and bias investigation	Untracked merging or silent dataset replacement	Assuming documented provenance ensures data quality	Independent lineage audit	Immutable version references and change logs
Code availability	Permit inspection and rerun	Model development and benchmarking	Executable code matched to manuscript version	Independent scrutiny and discrepancy detection	Incomplete scripts or missing dependencies	Treating repository presence as reproducibility	Fresh-environment execution	Release versioned, documented computational workflow
Environment capture	Preserve dependency and execution context	Deep-learning pipelines	Locked dependencies, containers or equivalent environment records	Reduce dependency-related replication failure	Hardware or external service mismatch	Assuming environment capture reproduces inaccessible data	Independent environment reconstruction	Archive environment specification with workflow version
Version control	Track evolution of code, data interfaces, and models	Iterative R&D workflows	Stable identifiers and disciplined change practice	Trace changes linked to results	Untracked manual edits	Assuming latest code produced published result	Commit/version linkage to outputs	Require release tags and result provenance
Random-seed reporting	Expose stochastic conditions	Neural networks; stochastic optimization	Seed control plus reporting of nondeterminism	Better interpretation of run-to-run variation	Hidden nondeterministic operations	Assuming one fixed seed establishes robustness	Repeated-run assessment where relevant	Record seeds and nondeterminism sources
Hyperparameter transparency	Make tuning choices inspectable	Model selection and benchmark comparison	Full search logic and selection criterion	Detect overfitting to benchmark or hidden tuning	Selective reporting	Treating final hyperparameters as full tuning evidence	Reconstruct selection procedure	Preserve search configuration and decision history

ISSN (Online) 2249-6084 (Print) 2250-1029

Preprocessing transparency	Reconstruct transformations before modeling	Molecular standardization, descriptor generation, multimodal integration	Ordered, versioned preprocessing steps	Detect leakage and transformation-induced bias	Global preprocessing using held-out information	Assuming preprocessing is scientifically neutral	Partition-aware pipeline audit	Fit data-dependent preprocessing only within appropriate training boundaries
Workflow provenance	Link inputs, code, environment, decisions, and outputs	End-to-end computational discovery workflows	Persistent identifiers and traceable execution history	Independent verification and auditability	Fragmented or unverifiable lineage	Assuming documentation volume equals traceability	End-to-end reconstruction test	Maintain machine-readable and human-readable provenance
Benchmark documentation	Clarify dataset, split, reuse, and evaluation assumptions	Comparative molecular ML	Transparent benchmark history and intended claim	Better interpretation of score meaning	Repeated tuning and contamination	Assuming standard benchmark equals unbiased benchmark	Untouched challenge set and leakage review	Record prior benchmark exposure and selection decisions

Validation and regulatory confidence

Validation in computational drug discovery should be understood as a layered challenge to a model claim rather than a single event completed when training performance, cross-validation, or a held-out score is reported. Training performance primarily describes optimization against observed data; internal validation evaluates behavior under a specified resampling or holdout design; scaffold-aware and temporal designs probe particular forms of structural or chronological separation; external validation challenges the model with data generated beyond the development set; prospective validation tests performance on future-facing cases; and experimental validation subjects selected computational outputs to physical or biological evidence. These layers differ in what they can establish, and none automatically proves universal generalizability. The same principle applies to uncertainty quantification. Comparative evaluation of scalable uncertainty-estimation methods for molecular property prediction indicates that uncertainty behavior is method- and context-dependent, so the presence of an uncertainty value should not be interpreted as evidence that confidence is appropriately calibrated for the decision at hand [20]. Validation confidence therefore depends on whether the evidentiary challenge matches the intended claim, anticipated distribution, chemical-space boundary, and consequence of error.

Internal validation can be informative while still leaving substantial uncertainty about transportability. Cross-validation and random holdouts may estimate performance under conditions resembling the sampled dataset, but they can overstate confidence when molecular redundancy, scaffold relationships, repeated assay structures, or preprocessing dependencies make evaluation partitions insufficiently distinct. Scaffold-aware validation is often used to increase structural separation, yet the choice of scaffold definition and the intended generalization target remain consequential; temporal validation may better approximate future-facing use when data-generating processes are sufficiently stable, but it can also expose drift in assays, chemistry programs, or organizational practice. Neural-network uncertainty studies in molecular property prediction likewise show that uncertainty quality varies across datasets, tasks, architectures, and estimation methods [21]. The relevant mechanism is therefore not "validation performed" or "uncertainty displayed," but exposure of the model to a challenge capable of revealing unsupported confidence under conditions that matter for use.

Prospective validation raises the evidentiary bar because the challenge is generated after model development rather than reconstructed entirely from historical data. In industrial ADME prediction, prospective evaluation has been used to test machine-learning algorithms against future cases in an operational setting, providing a more decision-relevant assessment than retrospective performance alone [22]. The trust-building mechanism is future-facing exposure: assumptions embedded in preprocessing, model selection, endpoint relationships, and chemical-space coverage are confronted by cases not available during development. Even so, prospective validation should not be equated with universal transportability. A prospective study may remain specific to one endpoint, organization, assay process, chemical program, or period. Confidence therefore increases in a bounded manner when the prospective context resembles intended use and when deviations, missingness, model updates, and selection processes are transparently documented.

Experimental validation provides another form of external challenge by testing computationally prioritized hypotheses against physical or biological evidence. Deep-learning-guided antibiotic discovery illustrates how computational prioritization can be connected to empirical testing, changing the evidentiary status of selected predictions without demonstrating that the model is valid for every compound class or discovery problem [23]. Similarly, deep-learning-enabled identification of DDR1 kinase inhibitors linked computational design to synthesis and experimental assessment, offering stronger evidence than an entirely in silico claim while remaining bounded to the studied workflow and target context [24]. These examples clarify an important distinction: prospective or experimental confirmation can strengthen translational confidence, but it does not establish clinical utility, regulatory acceptance, or universal model validity. Selection effects, narrow confirmation sets, assay choices, and incomplete replication may still constrain what can reasonably be inferred.

Validation should also expose chemically consequential failure modes that aggregate metrics can conceal. Data leakage may arise when information from held-out cases influences feature construction, preprocessing, target definition, or tuning; scaffold leakage may occur when structural relationships allow an apparently separated test set to remain easier than the intended extrapolation task; temporal leakage may introduce future information into historical prediction; and repeated benchmark use may convert an ostensibly untouched test set into an indirectly optimized resource. Distribution shift may additionally arise from changes in chemical series, assay practice, target biology, measurement technology, institutional context, or sampling policy. Activity cliffs are particularly important because small structural modifications can correspond to large property changes, creating local discontinuities that challenge models whose aggregate performance appears strong. Empirical work using activity cliffs to expose molecular machine-learning limitations demonstrates that apparently competitive models can remain brittle in chemically important local regions [25]. Validation should therefore include explicit scrutiny of chemical-space shift, local discontinuity, and applicability boundaries rather than rely exclusively on global averages.

Regulatory confidence sits downstream of, but is not identical to, scientific validation. In drug development, confidence may be strengthened when model purpose, data provenance, intended use, validation design, uncertainty, human responsibilities, version history, and change controls are sufficiently traceable to permit independent scrutiny. Analysis of AI and machine-learning applications in regulatory submissions shows that regulatory use is context-specific and embedded in particular development purposes rather than governed by a single threshold of technical performance [26]. Regulatory confidence therefore depends on fit-for-purpose evidence, auditability, risk proportionality, evidence provenance, model-update control, and clarity about how human judgment interacts with automated output. Regulatory confidence is not regulatory approval, and regulatory readiness is not clinical utility. **Table 3** synthesizes the major validation layers and regulatory-confidence mechanisms relevant to AI-enabled computational drug discovery.

Table 3. Validation and Regulatory Confidence in AI-Enabled Computational Drug Discovery: Evidence Layers, Contextual Preconditions, Confidence Mechanisms, Failure Risks, Documentation Needs, and Implementation Implications

Validation or confidence layer	Primary objective	Required context	Mechanism of confidence generation	Evidence produced	Major validity threat	Documentation requirement	Regulatory relevance	Implementation implication
Training performance	Determine whether the model can optimize against development data	Defined task, target, loss, and training process	Demonstration of fit to observed development data	Optimization and apparent learning behavior	Overfitting; leakage; poor target validity	Training data version, optimization logic, stopping criteria, model version	Low when isolated from independent validation	Necessary for development but insufficient for trust decisions
Internal holdout validation	Estimate behavior on held-out observations from the development data context	Partition created without information leakage	Separation of fitting and evaluation observations	Within-context predictive evidence	Hidden dependence; preprocessing leakage; repeated test use	Split construction, preprocessing boundaries, tuning procedure	Limited without evidence of intended-use relevance	Appropriate for development-stage comparison when claims remain bounded

Cross-validation	Assess variability across repeated internal partitions	Valid resampling unit and independence assumptions	Repeated challenge across development-data partitions	Distribution of within-context evaluation results	Related entities across folds; invalid resampling unit	Fold logic, grouping constraints, repeated-use history	Supportive but not evidence of external transportability	Useful for model selection when dependency structure is respected
Scaffold-aware validation	Challenge structural generalization beyond closely related chemistry	Defensible scaffold definition and intended chemical extrapolation claim	Reduction of direct structural similarity across partitions	Evidence of performance under stronger chemical separation	Scaffold definition may not match real deployment shift	Scaffold algorithm, partition statistics, chemical-space comparison	Relevant when intended use involves new chemical series	Should complement, not automatically replace, other validation designs
Temporal validation	Test future-facing behavior across chronological separation	Reliable timestamps and appropriate historical ordering	Exposure to later data unavailable during model development	Evidence under temporal evolution	Temporal leakage; assay/process drift; changing selection policies	Cutoff logic, data availability dates, evolving assay context	Stronger when future operational use is intended	Supports reassessment of drift and model-update needs
External validation	Challenge the model in a genuinely distinct data context	Independent or meaningfully different source, process, chemistry, institution, or population	Exposure to contextual differences not represented in development	Bounded generalization evidence	Nominal externality without substantive independence	Source provenance, overlap analysis, contextual comparison	Important for fit-for-purpose confidence	Externality must be demonstrated rather than assumed
Cross-institutional validation	Test portability across organizations or laboratories	Distinct workflows, infrastructure, data-generating practices	Institutional challenge of hidden local assumptions	Evidence about portability across sites	Harmonization artifacts; undisclosed overlap; local protocol differences	Site-specific provenance, harmonization steps, protocol differences	Potentially important for multi-site implementation	Reveals organizational dependencies requiring adaptation
Prospective validation	Assess performance on future cases after model development	Predefined workflow, future data, controlled model version	Future-facing exposure and reduced retrospective selection freedom	Evidence closer to intended operational use	Selective case inclusion; model changes during evaluation; drift	Prospective protocol, model lock or change history, inclusion rules	Strong when aligned with intended use	Can justify bounded operational confidence, not universal utility
Experimental validation	Challenge computational predictions with physical or biological evidence	Appropriate assays, controls, scientifically valid follow-up	Empirical falsification or confirmation of selected predictions	Experimental evidence for tested hypotheses	Selective testing; assay limitations; narrow confirmation set	Selection rationale, assay protocol, controls, provenance	Relevant when computational output influences development decisions	Strengthens translational evidence but does not prove broad model validity
Uncertainty calibration	Determine whether confidence estimates correspond meaningfully to error behavior	Defined prediction context and calibration criterion	Alignment of confidence with observed predictive reliability	Calibration evidence within tested conditions	Miscalibration under shift; inappropriate uncertainty proxy	Method, calibration set, metric, recalibration history	Relevant when uncertainty affects escalation or risk control	Uncertainty should change action only when validated for that use
Applicability-domain assessment	Identify where model support is adequate or weak	Explicit representation of	Recognition of unsupported or weakly	Boundary evidence	Incomplete domain characterization; novel shift	Domain definition, similarity/suppo	Supports risk-proportionate reliance	Use boundary information to trigger review,

ISSN (Online) 2249-6084 (Print) 2250-1029

		chemical/data support	supported predictions			rt measure, update logic		abstention, or further testing
Out-of-distribution assessment	Detect departures from development-data context	Meaningful reference distribution and shift criterion	Shift awareness before uncritical reliance	Evidence of potential context mismatch	Undetected novel shift; detector failure	Reference distribution, detection method, alert threshold rationale	Relevant to monitoring and model change	Should trigger investigation rather than automatic certainty
Traceability	Reconstruct how data, model, decisions, and outputs are connected	Stable identifiers and end-to-end provenance	Auditability and evidence reconstruction	Inspectable lineage	Fragmented records; untracked manual changes	Data lineage, code/model version, execution history	High for accountable use	Requires operational provenance rather than narrative description alone
Intended-use specification	Bound claims to a defined decision purpose	Explicit user, task, setting, consequence, and action	Alignment of evidence with decision purpose	Fit-for-purpose evidence statement	Scope expansion after validation	Intended users, decision role, exclusions, escalation pathways	Central to risk-proportionate evaluation	Prevents unsupported reuse of a model in higher-risk settings
Human oversight	Permit expert challenge, override, and escalation	Competent users with authority and adequate information	Human error interception and contextual judgment	Evidence of controlled decision process	Automation bias; unclear accountability; nominal oversight	Roles, authority, escalation and override procedures	Relevant when human review is part of the control strategy	Oversight must be functional, not symbolic
Change control	Preserve confidence when data, code, model, or use changes	Versioned lifecycle and material-change criteria	Reassessment of evidence after change	Documented continuity or revalidation decision	Silent model drift; uncontrolled updates	Change log, impact assessment, revalidation rationale	High for lifecycle confidence	Material changes should trigger proportionate reassessment
Monitoring	Detect degradation or contextual change after implementation	Observable inputs, outputs, process indicators, ownership	Continuing challenge of assumptions	Evidence of drift, failure signals, or stability	Unobservable outcomes; weak response process	Monitoring plan, ownership, alert and response logic	Relevant to sustained confidence	Monitoring without corrective authority is insufficient
Risk-proportionate governance	Match scrutiny to consequence, autonomy, uncertainty, and evidence maturity	Defined risk framework and accountable ownership	Allocation of stronger controls to more consequential decisions	Governance rationale and control evidence	Checklist compliance detached from scientific risk	Risk classification, control rationale, review record	Central to regulatory-confidence formation	Governance intensity should increase with potential harm and evidentiary uncertainty

Context–mechanism–outcome synthesis

The realist synthesis indicates that trustworthy AI cannot be explained by adding independent technical virtues to a model; rather, confidence emerges when contexts activate mechanisms that expose claims to meaningful challenge. Consider the explainability–scrutiny configuration. In a context of adequate data provenance, chemically meaningful representation, competent domain expertise, and explanation methods whose behavior is sufficiently assessed, an explanation can trigger scrutiny, error detection, and boundary awareness. The outcome may be more credible prioritization because the model becomes easier to challenge rather than merely easier to describe. The same logic extends to uncertainty-aware decision processes: drug-design scholarship on uncertainty quantification emphasizes that uncertainty becomes useful when it informs how decisions are prioritized, deferred, escalated, or experimentally challenged [27]. The mechanism is therefore not the numerical production of an uncertainty estimate but a change in reliance under conditions where the estimate is sufficiently credible for the relevant decision.

A contrasting configuration explains explanation overtrust. In contexts where post hoc outputs are visually persuasive,

ISSN (Online) 2249-6084 (Print) 2250-1029

stakeholders expect chemically intuitive narratives, fidelity is untested, or expertise is insufficient to interrogate the explanation, apparent transparency may activate acceptance rather than scrutiny. The outcome is false reassurance: an explanation that resembles domain knowledge can increase subjective confidence while leaving predictive validity, causality, and transportability unresolved. Chemical-space support modifies this pathway because even a seemingly coherent explanation may accompany a prediction outside the region adequately represented by development data. Work examining uncertainty and trust in AI-enabled drug discovery highlights the importance of applicability-domain reasoning and recognition of unsupported extrapolation [28]. Thus, the refined configuration is conditional: explainability may support trust when it enables challenge within a defensible domain, but it may undermine trust when plausibility substitutes for fidelity or when a confident narrative conceals weak chemical support.

The reproducibility–independent-verification configuration is similarly contingent. When data versions, preprocessing, code, dependencies, randomness, hyperparameters, and execution history are reconstructable, independent investigators can rerun the workflow, locate discrepancies, and identify boundary conditions. The resulting outcome is stronger scientific confidence because the claim becomes challengeable. When the same resources are absent, workflow opacity inhibits reconstruction and replication failure reduces confidence. Yet reproducibility alone remains insufficient: a fully reconstructable workflow can consistently reproduce leakage, biased labels, or an invalid endpoint. External challenge becomes important because it tests whether model claims survive conditions distinct from development. Translational AI literature emphasizes that representative local evaluation, distribution shift, workflow conditions, and deployment context shape whether apparent performance survives transfer [29]. The relevant mechanism is not an "external" label but exposure to differences capable of falsifying hidden assumptions.

A fourth configuration concerns leakage, memorization, and generalization. In contexts where related compounds cross evaluation boundaries, preprocessing incorporates held-out information, benchmark exposure influences repeated tuning, or endpoint structure is unusually redundant, the evaluation process may activate a shortcut mechanism: the model exploits relationships unavailable or less prevalent in intended future use. The outcome is inflated confidence rather than genuine transportability. This pathway can persist even when the workflow is reproducible and the model architecture is sophisticated. Comparative analysis of learned molecular representations across public and proprietary property-prediction tasks demonstrates that model behavior varies with dataset and endpoint context, arguing against architecture-level claims of universal superiority [30]. Consequently, generalization confidence should remain bounded by the exact challenge survived, including chemical-space distinctness, temporal separation, assay conditions, and the relevance of the external context to the intended decision.

The refined realist programme theory therefore proposes that trustworthy AI emerges when context-appropriate mechanisms are activated, independently challenged, and supported by validation and governance proportionate to decision risk. High-quality provenance and relevant chemical-space coverage can enable scrutiny and uncertainty interpretation; transparent workflows can enable independent reconstruction; leakage-aware evaluation can prevent some forms of performance inflation; external, prospective, or experimental challenge can test whether retrospective relationships persist; empowered human oversight can intercept errors; and traceable governance can preserve accountability across change. These mechanisms do not simply accumulate. They can correct one another: explanation may identify suspicious dependencies that trigger additional validation; uncertainty may prompt abstention or experimental confirmation; independent replication may reveal hidden preprocessing; and monitoring may expose a context shift requiring redesign. Trustworthy outcomes are therefore bounded and revisable rather than permanently conferred.

Figure 2 presents the refined realist programme theory explaining how trustworthy AI outcomes emerge from interactions among drug-discovery contexts, confidence-generating mechanisms, validation pathways, governance conditions, and implementation outcomes.

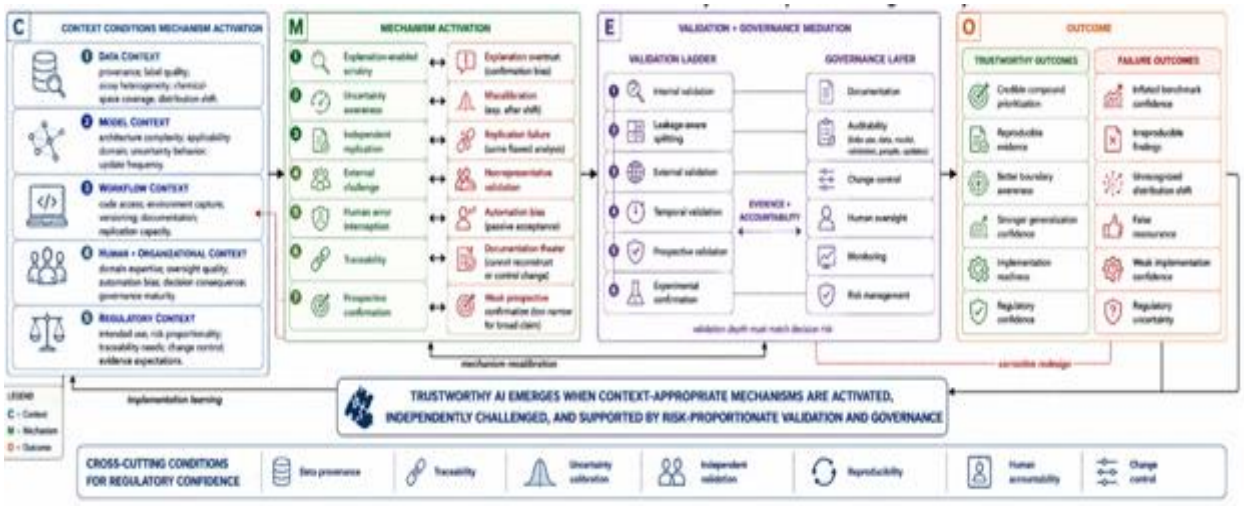

Figure 2. Context–Mechanism–Outcome Framework for Trustworthy AI in Computational Drug Discovery

Alt-text description

A wide realist framework organizes trustworthy AI into context, mechanism, and outcome layers. Contexts include data provenance, chemical-space coverage, assay quality, workflow maturity, stakeholder expertise, decision risk, and regulatory environment. Mechanisms include explanation-enabled scrutiny, uncertainty awareness, replication, external validation, human oversight, traceability, and prospective confirmation. Positive outcomes include credible prioritization, reproducibility, generalization confidence, implementation readiness, and regulatory confidence. Failure pathways show leakage, distribution shift, explanation overtrust, weak governance, and false confidence.

Implementation conditions

Operational trustworthy AI requires organizational conditions that convert abstract principles into repeatable practices across the model lifecycle. Data governance must preserve source provenance, transformation history, access controls, endpoint definitions, and responsibility for correction; workflow maturity must connect code, environments, model versions, preprocessing, validation artifacts, and decision records; multidisciplinary expertise must bring computational scientists into continuing interaction with medicinal chemists, biologists, toxicologists, quality specialists, and decision owners. Responsible-machine-learning guidance in high-stakes health contexts emphasizes lifecycle thinking, stakeholder engagement, careful problem formulation, and evaluation aligned with real use rather than isolated model construction [31]. For computational drug discovery, the transferable mechanism is accountable challenge: implementation becomes more trustworthy when domain experts can question assumptions, model developers can reconstruct evidence, and responsible decision-makers have authority to escalate, defer, override, or request additional validation.

Implementation must also recognize that regulatory confidence is institutionally contextual. Comparative analysis of AI- and machine-learning-based medical-device approvals across the United States and Europe demonstrates that regulatory pathways, product categories, and evidentiary environments differ rather than forming one universal model of acceptance [32]. Although medical-device evidence should not be treated as direct proof of drug-development requirements, the mechanism-level implication is relevant: regulatory confidence is shaped by intended use, jurisdiction, risk, evidence provenance, and the capacity to inspect how a system changes. Computational drug-discovery teams should therefore avoid treating "regulatory readiness" as a generic technical maturity score. Regulatory strategy should begin by specifying the model's role, consequence, degree of autonomy, evidence dependency, human control, and the conditions under which the system may be updated or repurposed.

Within pharmaceutical organizations, governance maturity becomes consequential when it is operational rather than aspirational. Pharmaceutical AI governance literature identifies the importance of accountability, data governance, transparency, cross-functional expertise, monitoring, and alignment between innovation and responsible control [33]. In practice, these conditions require named ownership, model documentation, code and environment management, validation planning, uncertainty communication, independent review, decision escalation, change control, and post-update reassessment. A material change in training data, endpoint definition, representation, architecture, software dependency, deployment context, or intended use should trigger proportionate review rather than silent continuation under

historical evidence. Governance failure occurs when records are incomplete, ownership is fragmented, model updates are disconnected from validation, or incentives reward rapid benchmark improvement while discouraging negative findings.

Governance intensity should be proportionate to decision consequence, uncertainty, model autonomy, evidence maturity, and potential harm. Early exploratory screening may tolerate greater uncertainty when model outputs are inexpensive hypotheses subjected to subsequent scientific filtering. Compound prioritization may require stronger chemical-space analysis, uncertainty communication, and documented human challenge because model rankings can redirect experimental resources. Lead-optimization support may justify more explicit provenance, local validity assessment, and change control as decisions become more costly and interconnected. Toxicity assessment and other high-consequence translational decisions demand stronger independence of validation, clearer applicability boundaries, robust escalation procedures, and tighter control of model updates. The realist implication is that one governance checklist cannot produce trustworthiness across all these settings: the same mechanism may be adequate in a low-consequence exploratory context and inadequate when an output substantially influences a high-risk development decision.

Future implementation priorities should therefore focus on infrastructures and incentives that make models easier to challenge. Standardized reporting should expose task formulation, dataset provenance, preprocessing, split logic, benchmark reuse, tuning decisions, and uncertainty assumptions; benchmark hygiene should address leakage, related entities, repeated test exposure, and inappropriate generalization claims; temporal and genuinely external validation should be used when they correspond to intended deployment; prospective or experimental challenge should be prioritized for consequential claims; uncertainty calibration should be connected to explicit actions rather than decorative confidence values; model-change governance should link material updates to reassessment; and auditability should preserve the chain from data to decision. Cross-disciplinary training is equally important because computational competence alone does not ensure chemically, biologically, organizationally, or regulatorily appropriate interpretation. Successful implementation is most likely when transparent provenance, independent challenge, human accountability, and risk-proportionate governance operate together; failed implementation and regulatory uncertainty remain likely when organizational incentives reward apparent performance while obscuring limitations, negative cases, and the cost of being confidently wrong.

CONCLUSION

This realist review reframes trustworthy AI in computational drug discovery as a conditional outcome emerging from interactions among context, mechanism, validation, and governance rather than as an intrinsic property conferred by accuracy, explainability, reproducibility, validation, or regulation alone. Explainability may support scrutiny, debugging, scientific plausibility assessment, and human challenge when explanations are sufficiently faithful, stable, domain-relevant, and competently interpreted, but it may produce overtrust when persuasive plausibility substitutes for evidence. Reproducibility can enable independent verification when data, code, environments, randomness, preprocessing, and workflow provenance are reconstructable, yet reproducible error remains error. Validation becomes confidence-generating when the challenge is distinct, relevant, leakage-aware, and proportionate to the intended claim; uncertainty awareness becomes useful when estimates are adequately evaluated and alter decisions appropriately; and regulatory confidence depends on traceability, intended-use clarity, human accountability, change control, and risk-proportionate evidence rather than documentation alone. The refined programme theory therefore holds that trustworthy outcomes emerge when appropriate contexts activate credible mechanisms, evidence is independently challenged, provenance remains transparent, uncertainty is acknowledged, humans retain meaningful responsibility, and governance intensity reflects decision consequence and evidentiary maturity. Trust should consequently remain bounded, revisable, and open to falsification as chemical space, workflows, models, organizations, and regulatory expectations change.

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. 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
3. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA, 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
4. 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
5. 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
6. 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
7. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-50. doi:10.1016/S2589-7500(21)00208-9
8. 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
9. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
10. de Weger EJ, van Vooren NJE, Wong G, Dalkin S, Marchal B, Drewes HW, et al. What's in a realist configuration? Deciding which causal configurations to use, how, and why. Int J Qual Methods. 2020;19:1609406920938577. doi:10.1177/1609406920938577
11. Greenhalgh J, Manzano A. Understanding 'context' in realist evaluation and synthesis. Int J Soc Res Methodol. 2022;25(5):583-95. doi:10.1080/13645579.2021.1918484
12. King H, Wright J, Treanor D, Williams B, Randell R. What works where and how for uptake and impact of artificial intelligence in pathology: Review of theories for a realist evaluation. J Med Internet Res. 2023;25:e38039. doi:10.2196/38039
13. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15. doi:10.1038/s42256-019-0048-x
14. McCloskey K, Taly A, Monti F, Brenner MP, Colwell LJ. Using attribution to decode binding mechanism in neural network models for chemistry. Proc Natl Acad Sci U S A. 2019;116(24):11624-9. doi:10.1073/pnas.1820657116
15. Wellawatte GP, Seshadri A, White AD. Model agnostic generation of counterfactual explanations for molecules. Chem Sci. 2022;13(13):3697-705. doi:10.1039/D1SC05259D
16. Rao J, Zheng S, Lu Y, Yang Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. Patterns (N Y). 2022;3(12):100628. doi:10.1016/j.patter.2022.100628
17. Amara K, Rodríguez-Pérez R, Jiménez-Luna J. Explaining compound activity predictions with a substructure-aware loss for graph neural networks. J Cheminform. 2023;15(1):67. doi:10.1186/s13321-023-00733-9
18. 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
19. 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
20. 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
21. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502
22. Fang C, Wang Y, Grater R, Kapadnis S, Black C, Trapa P, et al. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. J Chem Inf Model. 2023;63(11):3263-74. doi:10.1021/acs.jcim.3c00160
23. 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
24. 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

25. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073
26. Liu Q, Huang R, Hsieh J, Zhu H, Tiwari M, Liu GS, 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
27. Mervin LH, Johansson S, Semenova E, Giblin KA, Engkvist O. Uncertainty quantification in drug design. Drug Discov Today. 2021;26(2):474-89. doi:10.1016/j.drudis.2020.11.027
28. Yu J, Wang D, Zheng M. Uncertainty quantification: Can we trust artificial intelligence in drug discovery? iScience. 2022;25(8):104814. doi:10.1016/j.isci.2022.104814
29. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
30. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88. doi:10.1021/acs.jcim.9b00237
31. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6
32. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. Lancet Digit Health. 2021;3(3):e195-203. doi:10.1016/S2589-7500(20)30292-2
33. Pasas-Farmer S, Jain R. From discovery to delivery: Governance of AI in the pharmaceutical industry. Green Anal Chem. 2025;13:100268. doi:10.1016/j.greeac.2025.100268