From Chemical Structure to Biological Mechanism

The causal reasoning challenge in AI-enabled drug discovery

Abdelrahman Alabed BSc

Head of AI Strategy & Scientific Intelligence - aveiaa

May 2026

Abstract

Artificial intelligence (AI) has become an essential pillar in drug discovery processes, including compound design, virtual screening, ADMET prediction, target prioritization, and clinical development planning. However, making precise predictions about whether a molecule will be effective and safe in the human body remains a significant challenge, considering that the majority of molecules optimized by predictive models or shown to be active in biochemical assays still fail in the later clinical phases.

The high failure rate can be attributed to the complexity of the chemical and biological context. This complexity exceeds what can be integrated into any prediction model. to address this issue, causal reasoning offers frameworks for connecting chemistry, disease biology, and clinical translation. Nevertheless, a fundamental limitation of current causal reasoning frameworks is their difficulty in capturing all necessary evidence and connecting it through causal relationships without missing important drivers of program success. The "driver" here is not necessarily more data; it is the specific evidence or causal rule needed to connect a compound's behavior in a model system to its likely behavior in humans.

In this article, we focus on the challenges facing causal reasoning in drug discovery, including data availability and validity, causal variable definition, evidence validity, chemistry–biology translation, context preservation, technical artefacts, uncertainty, and acceptance by scientific and regulatory stakeholders. The main argument is that industry-ready causal reasoning systems must distinguish association from intervention, preserve provenance, expose uncertainty, and support human scientific judgment rather than replace it.

The final section describes how new multi-domain causal engines such as aveiaa's Knowledge Vault and CMDN reasoning layer can help address causal reasoning challenges by providing practical solutions to the issues discussed in the paper. Overall, we aim at this article to provide clear guidance for industry practice on the development and use of future causal reasoning systems.

Keywords: causal reasoning; drug discovery; AI-enabled drug discovery; data validity; literature mining; knowledge graphs; translational science; explainable AI; aveiaa

1. Introduction: prediction is not the same as intervention

Artificial intelligence is accelerating drug discovery by improving prediction across chemical, biological, and translational tasks. Current machine learning tools can generate rapid predictions at many levels, from molecular properties and target interactions to pathway activity and clinical response. However, faster prediction does not by itself solve the central problem of drug discovery: what will happen if this molecule is synthesized, dosed, and introduced into a living human system?

While the associations observed by AI models provide a critical window into answering this question, it is essential to recognize that these associations are only part of the broader evidence needed to make such a decision. Models can learn that a molecular descriptor, substructure, gene-expression signature, or pathway label is associated with activity in a dataset. But these associations do not inherently mean that modifying the molecule will modify the disease outcome. Causal inference provides a framework for bridging this gap between observing a pattern and reasoning about the likely effect of an action or intervention [1,3,4].

For industry teams, this distinction is operational rather than theoretical. It affects target selection, indication expansion, biomarker strategy, combination design, safety assessment, and go/no-go decisions. A causal reasoning system that cannot distinguish strong intervention evidence from weak association evidence may increase confidence without increasing reliability.

In drug discovery, causality can be considered at two connected levels: within domains and across domains. Within the chemical domain, a structural change may alter physicochemical properties, binding affinity, or metabolic stability. Across domains, that same chemical change must influence exposure, tissue access, target engagement, and ultimately a disease-relevant biological state. Target engagement alone is not sufficient; it must propagate through the relevant pathway in a way that changes the phenotype of interest. This effect must also remain meaningful across model systems, species, and patient populations. If the causal chain breaks at any step, the development case may fail, even when earlier predictions were accurate within their own limited domains.

2. The chemistry-biology interface: why linear translation breaks down

Medicinal chemistry usually begins with local changes: a ring is replaced, polarity is adjusted, a basic center is removed, or a substituent is moved to another position. On the other hand, biology is wider and subject to diverse contributors such as target expression, pathway redundancy, compensatory signaling, cellular state, tissue context, metabolism, exposure, transport, protein binding, genetic background, and disease heterogeneity. Once a chemical change enters a biological system, its effect is shaped by the wider biological context. As a result, the same molecule can produce different outcomes in different settings.

For instance: BRAF inhibition in melanoma can produce strong clinical responses. In BRAF V600E-mutant colorectal cancer, however, BRAF blockade alone was limited by EGFR-mediated feedback reactivation of MAPK signalling [5]. Similarly, gefitinib is considered effective in EGFR-mutant non-small-cell lung cancer, but its activity is strongly dependent on tumor genotype; because there is resistance that commonly emerges through mechanisms such as the EGFR T790M mutation [6]. These bypass and resistance mechanisms show that biological systems can reroute, compensate, or escape even when the original target hypothesis is reasonable.

A similar problem appears in many computational drug discovery workflows, including QSAR, molecular docking, deep learning models, knowledge graphs, and digital twins. These tools have clearly improved the speed of screening, prioritization, and hypothesis generation [7-10]. However, in practice, a high-scoring prediction is rarely enough to justify a development decision by itself. The more important question is how that prediction fits into the causal evidence linking the molecule to the target, the target to the relevant pathway, and the pathway to a disease-related phenotype. This point is not meant to suggest that these computational approaches are weak or inefficient. Rather, their strongest value appears when they are used as part of a causal reasoning framework, where evidence from chemistry, biology, and translational science can be connected while still accounting for mechanism, biological context, uncertainty, and the feasibility of intervention.

3. Core challenges in causal reasoning across chemistry and biology

3.1 Causal variables are not handed to the model

Typically, classical causal inference starts after variables have already been defined. In drug discovery, the situation is different. The variables themselves must be derived from different types of chemical, biological, experimental and clinical evidence. This creates a critical risk because choosing wrong variables will make a causal analysis look mathematically clean while remaining biologically uninformative.

This is why learning causal representations is critical. The variables that matter in drug discovery are hidden within noisy assay readouts and indirect molecular measurements. Schölkopf et al. described the discovery of high-level causal variables from low-level observations as a central problem at the interface of AI and causality [2]. In the context of drug discovery, this means that a weak assay response may reflect poor target engagement, limited cellular exposure, pathway independence, or an artifact of the disease model. Treating all of these possibilities simply as “low activity” can make the analysis mathematically clean but biologically misleading. The real value is to learn representations that separate these possibilities and connect chemical changes to the mechanisms that determine efficacy, safety, and translation to humans.

3.2 Chemistry and biology speak different evidence languages

Chemical data starts with molecules. Biological data starts with genes. Moving forward, chemical information starts to describe structure, descriptors, docking poses, etc., while biological information starts to reference genes, proteins, pathways, etc. Chemistry doesn't translate one-to-one to biology.

Putting both data types in the same database or knowledge graph is therefore only a start. The interaction between a drug and a target can mean many things, such as measured binding, predicted binding, curated inhibition, genetic validation, clinical exposure, etc. grouping these edges into one type weakens causal reasoning, because causality in biology is usually conditional. The same target can be causal in one disease subtype and irrelevant in another. The same compound can look active in a screening model and fail in a patient-derived system. The same pathway can be protective in one tissue and pathological in another. Mutation status, tissue origin, species, cell state, co-treatment, dosing schedule and disease stage can all change the interpretation of an otherwise reasonable result.

This creates a problem for broad AI models. Confidence is not enough if the evidence does not apply to the case being evaluated. A causal system must therefore preserve the source, evidence type, direction, sign, experimental context, and biological context of each relationship.

3.3 Data availability and validity are part of the causal problem

Causal reasoning is highly dependent on data availability because missing data can change the structure of causality. Most published literature and data sources contain positive associations far more than failure data. In practice, this can lead to overly optimistic causal graphs. Additionally, the absence of context in a database means that the same variable could be reused in different contexts with the same impact on results.

Validity is another critical concern, especially when claims are mined from the literature. While this is usually useful, it can also result in incorrectly mined claims. The problem here is that an association between variables can be stated in a sentence without any evidence of successful intervention. For instance, studies may observe an effect in a cell line, but this effect may poorly represent patient populations. When such claims are fed into the causal graph, the system will start to reason from a warped map of biology.

This produces a causal-cycle problem when biased claims are mined as evidence, transformed into causal relations, used to justify a downstream phenotype, and then used again to propose the next experiment. At this

point, the error starts to become harder to detect because it has been embedded in a logical sequence. Therefore, without causal systems powered by strict validation layers, the output causal graphs will be misleading and irrelevant.

3.4 Technical artefacts can become false biology

High-throughput biology is vulnerable to batch effects, platform differences, reagent lots, sample handling, sequencing depth, image-acquisition settings and laboratory protocols. These sources of variation can produce signals that resemble biology. Conversely, aggressive correction can remove true patient-relevant heterogeneity. The risk is not only lower accuracy; the risk is a false mechanism.

Leek and colleagues showed how batch effects in high-throughput data can lead to incorrect conclusions when technical variation is correlated with the outcome of interest [13]. In causal reasoning, the same problem becomes more serious. If a technical artefact is encoded as a biological edge, the system can build a plausible but false mechanistic chain. Any serious causal platform must therefore store provenance, experimental design, processing history and measurement context alongside biological assertions.

3.5 Association, mechanism and causation are often collapsed

Biomedical language often mixes several kinds of claims. A paper may say that a target is associated with disease, a pathway is upregulated, a compound reduces a phenotype, a biomarker predicts response, or a genetic variant modifies risk. These statements do not carry the same causal weight. Some support plausibility. Some indicate perturbational evidence. A smaller subset supports intervention.

Pearl's causal diagram framework and subsequent work on causal discovery make clear that statistical relationships must be combined with assumptions and subject-matter knowledge before causal effects can be identified [3,4]. Drug discovery often violates the clean assumptions of textbook causal inference: data are sparse, selectively published, confounded by experimental design, shaped by missing negative evidence and affected by unmeasured variables. Collapsing evidence into a single confidence score can therefore produce an attractive but misleading story.

3.6 Target biology and compound biology must be separated

A biologically attractive target does not guarantee a developable drug. The compound brings its own causal profile: tissue distribution, permeability, metabolism, selectivity, off-target pharmacology, residence time, safety margin, formulation feasibility and exposure dynamics. Conversely, a compound may produce a phenotype through a mechanism unrelated to its intended target.

Drug discovery therefore requires at least two linked causal questions. First, does the target or pathway matter for the disease? Second, can a particular molecule perturb that target or pathway in the right way, at the right site and within a feasible therapeutic window? Network methods can help because biological effects rarely arise from one isolated target. Still, a network path is not proof of therapeutic causality. Direction, dose, sign, context and evidence quality remain decisive.

3.7 Translation requires evidence across layers

A development decision is stronger when independent evidence layers point in the same direction. Human genetics can improve confidence in target selection, but it does not solve compound feasibility. A cell phenotype can suggest pathway relevance, but it may not predict exposure, safety or patient response. A digital twin or programmable virtual human concept can support scenario testing, but it remains limited by the validity, coverage and context of its data and model assumptions [8,9].

Evidence-integration platforms such as Open Targets illustrate the value of combining genetics, expression, literature, perturbation and clinical information to support therapeutic hypotheses [11]. The next step is more decision-specific. Teams need to translate integrated evidence into an argument about a particular molecule, mechanism, patient context and next development action.

3.8 Applicability domains and uncertainty are still under-specified

Predictive models are reliable only within the conditions where their assumptions hold. In QSAR and toxicology, this principle is described as the applicability domain. OECD guidance on QSAR assessment emphasizes the need to evaluate models and predictions in relation to intended use, reliability and uncertainty [12]. The same logic applies to causal AI.

A useful causal output should not simply state that a molecule looks promising. It should say which evidence supports the conclusion, where the system is extrapolating, which assumptions are active, what evidence is missing, how conflicts were handled and which experiment would most efficiently reduce uncertainty. Without this discipline, AI-generated reasoning can sound precise while remaining fragile.

3.9 Scientific and regulatory acceptance require explanation

Drug development decisions affect patient safety, capital allocation, regulatory strategy and scientific credibility. A causal reasoning system cannot function as a black box. Chemists, biologists, translational scientists, clinicians, regulatory experts and investors may need different levels of explanation, but they all need traceability.

Explainable AI methods can help expose model behaviour, but explanations also have failure modes. Recent reviews highlight shortcut learning, dataset bias, unstable attributions and the risk that an explanation appears chemically plausible while reflecting a dataset artefact rather than true pharmacology [10]. This reinforces the need to connect explanation with evidence validation, mechanistic plausibility and experimental follow-up.

4. Toward validation-aware causal intelligence

The practical lesson is straightforward: a larger evidence graph is not necessarily a better causal model. Before a platform reasons across chemistry and biology, it must decide which claims are safe enough to use, what each claim means and where it applies. A validation-aware causal system should therefore combine evidence memory, causal structure, context gating, uncertainty management and human review.

4.1 Evidence memory rather than document storage

A scientific evidence memory should not store only documents. It should store claims as structured evidence objects. Each object should retain source, provenance, evidence type, method, context, directionality, sign, confidence, contradictions and missing assumptions. This allows a scientist to audit why a relation exists in the system and whether it should be used for a particular decision.

4.2 A causal validation layer

The validation layer is the point at which data quality and causal reasoning meet. A claim should be checked before it becomes part of a causal chain. Does the evidence support association or intervention? Is the direction known? Is the sign of effect coherent with upstream and downstream evidence? Is the study context compatible with the use case? Are there conflicting studies? Is the claim repeated across sources or merely duplicated from the same origin?

Causal validation is not the same as text validation. The question is not only whether a sentence was extracted correctly. The question is whether the extracted claim is usable for intervention-oriented reasoning.

A validated evidence object may still be uncertain, but its uncertainty should be visible and correctly located in the reasoning process.

4.3 Context gates and conflict preservation

Context gates prevent evidence from being applied where it does not belong. A pathway observation in a transformed cell line should not be generalized to a patient subgroup without qualification. An animal result should not be treated as human evidence. A genetic association should not be treated as compound feasibility. Context gates keep the system honest by limiting inference to the biological and experimental setting in which evidence has meaning.

Conflicts should also be preserved. In biomedical science, contradictory findings often reveal differences in assay system, disease subtype, dose, endpoint or patient population. Averaging conflicts into one score may remove the information needed for translation. A good causal system asks why evidence disagrees.

4.4 Uncertainty and next-experiment reasoning

Early discovery almost always operates under sparse evidence. The useful output is not certainty. It is a structured account of what is known, what is assumed, what is uncertain and what would most change the decision. This is where causal reasoning becomes operational: it can identify the missing link whose validation would most reduce development risk.

5. Aveiaa perspective: a measured implementation view

Aveiaa is positioned around the problem described in this article: how to turn scattered scientific evidence into causal reasoning that scientists can inspect and use. Knowledge Vault™, and CMDN™, are causal evidence repository and reasoning layer that operate on a validation powered chemistry-biology causal framework designed to identify the correct causal path being proposed, the proper evidence supporting each step, the evidence that contradicts it, the assumptions that are being made and the experiments that would reduce uncertainty.

For early discovery teams, this can improve go/no-go discussions before downstream work becomes expensive. For research institutions and biotechnology companies, it can support assay selection, grant applications, partnership discussions and prioritisation. For CROs and development groups, it can make assumptions and evidence gaps visible before a programme is advanced into costly validation work.

7. Concluding remarks

Causal reasoning across chemistry and biology is one of the central operational problems in modern drug discovery. The challenge is not simply to build more predictive models or larger databases. It is to connect chemical structure, compound behaviour, target engagement, biological mechanism, translational evidence and clinical relevance without losing context, validity or uncertainty.

Data availability and validity deserve special attention. Biased, incomplete or incorrectly mined evidence can mislead causal reasoning more severely than it misleads ordinary prediction, A validation layer that uses causal reasoning to test, qualify and contextualize evidence is essential to overcome these issues.

Aveiaa's Knowledge Vault, CMDN concepts are aligned with this need when they are understood as evidence-validation and reasoning infrastructure. Their scientific value preserve provenance, separate evidence types, gate context, manage uncertainty, expose conflicts and support human review. The goal is not to claim that a molecule will succeed. The goal is to make clearer why it might succeed, where it might fail and what should be tested next.

References

[1] Minikel EV, Painter JL, Dong CC, et al. Refining the impact of genetic evidence on clinical success. Nature. 2024;629:624-629. doi:10.1038/s41586-024-07316-0.

[1] Michoel T, Zhang JD. Causal inference in drug discovery and development. Drug Discovery Today. 2023;28(10):103737. doi:10.1016/j.drudis.2023.103737.

[2] Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y. Toward causal representation learning. Proceedings of the IEEE. 2021;109(5):612-634. doi:10.1109/JPROC.2021.3058954.

[3] Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669-688.

[4] Spirtes P, Zhang K. Causal discovery and inference: concepts and recent methodological advances. Applied Informatics. 2016;3:3. doi:10.1186/s40535-016-0018-x.

[5] Prahallad A, Sun C, Huang S, et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature. 2012;483:100-103. doi:10.1038/nature10868.

[6] Kobayashi, S. et al. EGFR Mutation and Resistance of Non–Small-Cell Lung Cancer to Gefitinib. New England Journal of Medicine 352, 786–792 (2005). This is a classic reference for acquired resistance to gefitinib involving the EGFR T790M mutation.

[7] Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019;18:463-477. doi:10.1038/s41573-019-0024-5.

[8] Wu Y, Bourne PE, Xie L. AI-powered programmable virtual humans toward human physiologically-based drug discovery. Drug Discovery Today. 2025;30(11):104497. doi:10.1016/j.drudis.2025.104497.

[9] Venkatapurapu SP, Clegg L, Nowojewski A, Kimko H, Olabode D, Sawant-Basak A, Vishwanathan K. Digital twins for accelerating drug discovery and development: opportunities and challenges. Drug Discovery Today. 2026;31(2):104617. doi:10.1016/j.drudis.2026.104617.

[10] Gangwal A, Lavecchia A. Explainable AI methods for drug discovery: a survey of interpretability, metrics and mechanistic insight. Computer Science Review. 2026;61:100943. doi:10.1016/j.cosrev.2026.100943.

[11] Ochoa D, Hercules A, Carmona M, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Research. 2023;51(D1):D1353-D1359. doi:10.1093/nar/gkac1046.

[12] OECD. (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions. OECD Series on Testing and Assessment, No. 386. Paris: OECD Publishing; 2023. doi:10.1787/d96118f6-en.

[13] Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. 2010;11:733-739. doi:10.1038/nrg2825.

[14] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics. 2006;7:119-129. doi:10.1038/nrg1768.

[15] Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012;483:531-533. doi:10.1038/483531a.

[16] Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005;2(8):e124. doi:10.1371/journal.pmed.0020124.