

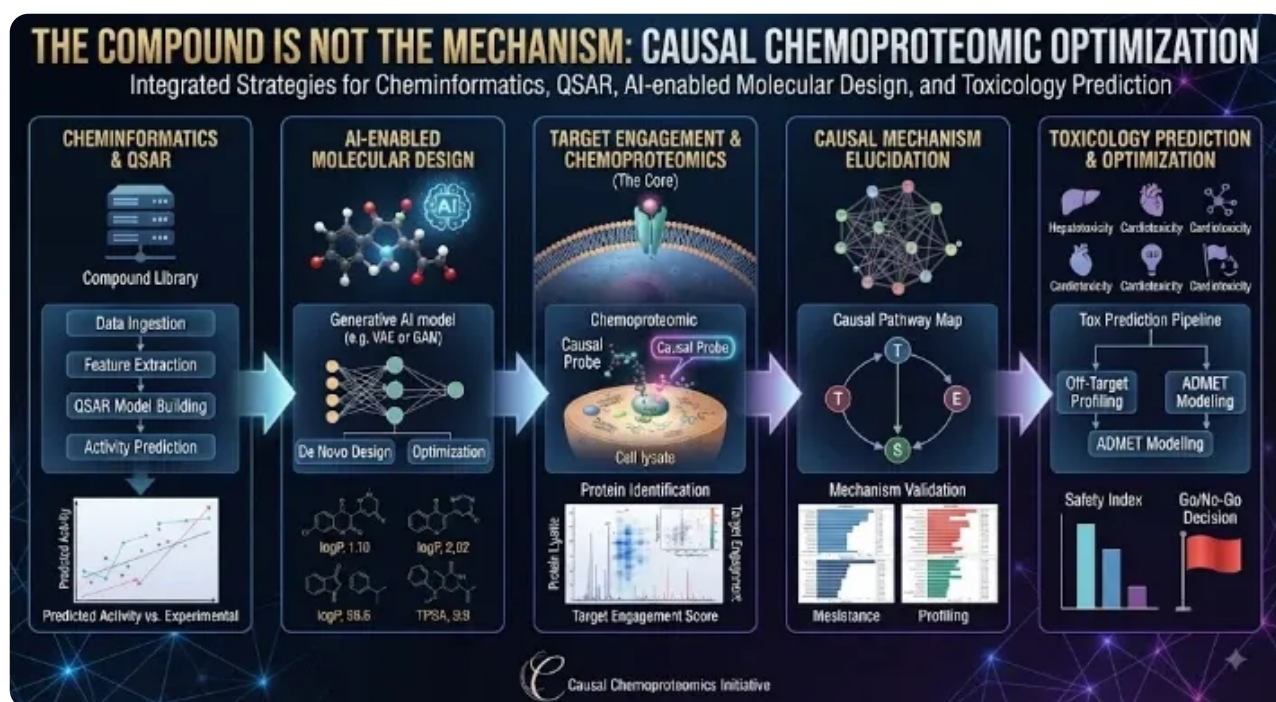
The Compound Is Not the Mechanism: Causal Chemoproteomic Optimization of Cheminformatics, QSAR, AI-enabled molecular design and toxicology prediction

Md. Humayan Kabir

Feb 23, 2025

Abstract

This article argues that QSAR and AI predictions of activity or toxicity should not be mistaken for biological mechanisms. Statistical models can correlate chemical structure with outcomes while missing the actual cellular interactions that cause therapeutic effects or adverse responses. Causal chemoproteomics adds measured protein engagement to computational descriptors, creating mechanistic labels that endpoint datasets cannot reliably infer. The article emphasizes scaffold-based and time-based validation because random splits can place close analogues in both training and test sets and exaggerate performance. Applicability domains and uncertainty calibration are used to flag novel chemistry. Training domain, descriptors, scaffold novelty, concentration, reactive groups, target abundance, metabolites, cell type, and assay endpoint are treated jointly. Failure routes include dataset bias, activity cliffs, hidden confounding, off-target binding, metabolic activation, domain shift, and false confidence outside training chemistry. Cellular concentration, protein binding, reactive metabolite exposure, tissue accumulation, and time-dependent engagement define exposure. Prospective compounds, orthogonal toxicology assays, target competition, genetic perturbation, metabolite testing, and independent chemical series are required to connect predicted risk with measured cellular cause.



The Compound Is Not the Mechanism highlights one of the main challenges in causal chemoproteomics for AI-driven molecular design and toxicology prediction, namely, the ability to predict activity/targets or toxicity using QSAR and machine learning while missing the actual cellular interactions. The more robust development approach is based on an actual quantifiable mechanism, not a single (wowza) assay result. With that change, chemistry, biology, exposure, resistance, and safety, become components of one translational argument, as does production.

Causal chemoproteomics provides measured protein engagement to computational descriptors to transform statistical associations into experimentally testable mechanisms with the scientific foundation. There is no single measurement that can be used to support the entire mechanistic claim, as each one addresses a separate question. It minimizes the potential for misinterpretation of potency, binding, formulation performance and/or model fit as therapeutic understanding.

A well-performing QSAR and machine-learning model typically performs best when it is only tested on random partitions of the data, as it is possible for close analogues to be in both the training and test sets. Validation using scaffolding or time results in a more challenging test, one more like discovery. Another layer of protection is provided when new chemistry is out of the training distribution through applicability domains and uncertainty calibration. Chemoproteomics then provides mechanistic labels which can only be guessed at with reliability from endpoint toxicity values by pure structure-property models.

Sometimes the design variables can start to influence each other and alter performance at the same time, necessitating joint treatment of the design variables. This optimization of one factor at a time can, in many cases, result in a negative impact on one or more other factors, particularly potency versus stability, selectivity versus delivery, or manufacturability. Therefore, multiparameter decision rules will give the candidates more realistic ranking than a single numerical endpoint.

Early failure routes measurement is also an important factor for mechanistic confidence. Key escape mechanisms include dataset bias, activity cliffs, hidden confounding, off-target binding, metabolic activation, domain shift and false confidence outside of training chemistry. These routes can be used as stress testing to turn them into design information to use in the next chemistry, formulation, dosing, or model cycle.

The link between nominal dose and the biological effect is the exposure. Relevant parameters include: cell concentration, protein binding, metabolic stability, reactive metabolite exposure, tissue accumulation and time-dependent target engagement. A simplified assay may score high on intrinsic activity but have a low score for exposure at the causal site, and thus may not be suitable for translation.

Interpretation of direct engagement or functional perturbation must be done at realistic concentrations. High dosages may lead to secondary responses to stress, off-target effects, or assay artifacts which mimic a desired mechanism. Concentration-response analysis in relation to exposure/occupancy is useful to determine the range in which the proposed causal process is in effect.

Bio-context of interpretation is also a factor in cheminformatics and computational toxicology programs that aim to prioritize synthesis or experimental safety testing. The dependency and effective exposure may vary depending on genotype, cell state, tissue architecture, immune tone, metabolic environment and previous treatments. Representative model panels are thus more reliable than multiple tests in a single system.

There is a need to measure resistance and adaptation in a prospective manner rather than to give them an ex post retroactive explanation. The mechanisms that survive repeated pressure or long term selection can be determined by serial passage, long term selection, genetic perturbation, repeated dosing, or environmental stress. A not as strong a candidate with a higher escape barrier can be more translationally valuable than a weak screening leader.

Assessment of safety is more informative if it is linked to mechanism. Proteome-wide

reactivity, mitochondrial stress, nuclear receptor signaling, ion-channel effects, genotoxicity, hepatic injury signals and off-target occupancy are all important considerations. The comparison of desired engagement versus unwanted biological responses allows discrimination between mechanism related pharmacology versus general cell injury or formulation related toxicity.

The manufacturing process cannot be a downstream logistics challenge as compound identity, purity, stereochemistry, batch consistency, solubility, assay preparation and reproducible standards impact the product that is presented to biological systems. Even if the product name remains the same, small changes to the process can result in changes to composition, physical state, level of impurities, stability, and/or dose reproducibility. Robustness of development is enhanced if attributes of the critical process are directly related to the mechanistic performance.

Applicability domain, uncertainty calibration, scaffold splits, external validation, explainability descriptors, counter-factual design and causal data fusion are computational tools that provide leverage. The value of prediction is greatest when there is some uncertainty and applicability limitations and particularly when the structures are new, unusual biological states and/or the datasets are sparse. Models that reveal uncertainty can facilitate experiment selection than models that give confident rankings but do not reveal places where uncertainty exists.

Going forward, prospective validation will need to be done with prospective compounds, orthogonal toxicity assays, target competition, genetic perturbation, metabolite testing and independent chemical series. Orthogonal tests are important because if there are repeated measurements from the same family of assays then the same bias can be reproduced without reinforcing causality. If there is agreement between independent methods, then the design choice to the biological consequence is strengthened; if there is a disagreement, it is an indication of assumptions that need to be modified.

Data should be integrated so that the traceability of the raw data to the development decision is maintained. Chemical structures, process conditions, assay metadata, exposures data, omics features, resistance data, and safety signals should all be made identifiable and versioned. This discipline helps to avoid later misinterpretations of the results that are related to conditions other than those which actually caused the result.

Translation should not be laboratory spectacle and the indicators of decision-making should reflect that as well. Score cards that are useful include reproducibility, activity normalised by the exposure, mechanistic confidence, resistance frequency, safety margin,

stability and manufacturing robustness. This will incentivize balanced candidates, as well as provide transparency in making compromises, that can be costly without committing to further development.

The negative results in this framework have unusually high values. The limits of a hypothesis are set by the failure to synthesize, inability to achieve the desired target engagement, or instability of the exposure, failure to rescue, resistance emergence, poor batch reproducibility, or loss of activity in realistic models. Keeping track of those failures helps with designing experiments that will lead to further discoveries and avoid reinvesting in appealing yet unsupported explanations.

The filter of reproducibility across laboratories and across batches is also a must. No matter how statistically significant a result is, a mechanism relying on just one operator, one preparation, one cell line or one modeling option is not free of fragility. Durable biology is determined by independent repetition under exposure matched conditions.

A down to earth perspective of clinical or deployment conditions is also essential in the process of translation. Roofing and administrative complexities as well as requirements for diagnosis, storage, monitoring, cost and treatment timing can make a technically advanced product attainable or not. The impact of mechanistic excellence comes to the fore when product design is applied in the context in which biological benefit can only be effected.

Temporal design is an additional layer due to variation in biological system response to short pulses, long exposure, washout and multiple dosing. Time-resolved measurements can distinguish between initial engagement and downstream adaptation, and can indicate if initial activity continues following the diminishing original signal. Furthermore, sampling schedules should match the expected kinetics of causal chemoproteomics for molecular design and prediction of toxicology (via AI) in order to avoid only one convenient endpoint.

A medicinal chemistry and formulation science integration development record helps in better communication throughout medicinal chemistry, formulation science, pharmacology, computation, manufacturing and clinical planning. The record should not be one definitive story, but rather provide a distinction between what is known, what seems to be likely, what is not known or certain, and what is the next experiment with the highest information value. This transparency helps iteration to be faster, and decreases the pressure to justify early assumptions when there is conflicting data.

Benchmarking against credible reference conditions helps to maintain the sense of novelty. New approach is determinative whether it is better to predict, explain, durable and/or more manufacturable than established compounds (known inactive), standard formulations, baseline models or accepted analytical means. The head-to-head comparison also reveals gains which are lost following normalization for exposure, assay sensitivity, batch quality or experimental complexity.

Predicted risk is more relatable when it is associated with measured cellular interactions and explicit uncertainty, making AI more reliable. It is not a winning or losing point of an assay, but the proposed mechanism that still makes sense when considering exposure, biological variation, resistance, safety, and production. That benchmark provides a more regimented transition from discovery signal to a therapeutic concept that will stand up to complexity in the real world.