Pharmacophore

ISSN-2229-5402

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

THE DIGITAL MOLECULAR TWIN AS A LIVING ACCOUNT OF STRUCTURE, DYNAMICS, BINDING, METABOLISM, TOXICITY, AND OFF-TARGET RISK

Laura Keller[1]*, Thomas Lehmann[1], Simon Brunner[2], Christoph Meier[3]

1. ***Department of Digital Molecular Twins and Living Accounts, Faculty of Pharmacy, ETH Zurich, Zurich, Switzerland.***
2. ***Department of Structure–Dynamics–Binding Integration, Faculty of Pharmacy, EPFL Lausanne, Lausanne, Switzerland.***
3. ***Department of Metabolism–Toxicity–Off-Target Risk Modeling, Faculty of Pharmacy, University of Bern, Bern, Switzerland.***

ARTICLE INFO

Received:
16 August 2025
Received in revised form:
04 December 2025
Accepted:
08 December 2025
Available online:
28 December 2025

Keywords: Digital molecular twin, Molecular state representation, Molecular dynamics, Binding free energy, Predictive ADMET, Off-target risk

ABSTRACT

Computational drug discovery commonly represents a molecule through a fixed chemical graph, descriptor vector, three-dimensional conformer, or learned embedding and then evaluates that representation against one or more isolated prediction tasks. Although such representations are useful, they cannot by themselves maintain an evolving account of how molecular identity, conformational behavior, target engagement, metabolism, toxicity, and off-target interactions change with biological context and accumulating evidence. This article proposes the digital molecular twin as a versioned, molecule-specific computational architecture rather than as a single predictive model. The proposed construct comprises linked state variables for molecular identity, structural ensembles, dynamics, binding, disposition, metabolic transformation, toxicity, and off-target risk; an evidence ledger that preserves source conditions and provenance; update operators that integrate simulation and experimental observations without silently overwriting prior states; a counterfactual query layer for examining possible lead-optimization changes; and an uncertainty layer that records model, data, sampling, and context limitations. The central contribution is an architectural distinction between predicting an endpoint and maintaining a living, auditable account of the evidence relevant to a molecule. The framework further separates association from mechanism, prediction from experimental confirmation, benchmark performance from pharmaceutical usefulness, and local technical validity from decision fitness. It is presented as an original conceptual and methodological synthesis that requires prospective evaluation, domain-specific calibration, and human oversight. It does not establish universal state definitions, validated causal inference, clinical utility, regulatory acceptability, autonomous optimization, or operational readiness. By organizing heterogeneous computational and experimental evidence around explicit states, versions, uncertainties, and decision boundaries, the digital molecular twin may provide a disciplined basis for developing more traceable and scientifically bounded molecular reasoning systems.

To Cite This Article: Keller L, Lehmann T, Brunner S, Meier C. The Digital Molecular Twin as a Living Account of Structure, Dynamics, Binding, Metabolism, Toxicity, and Off-Target Risk. Pharmacophore. 2025;16(6):99-108. https://doi.org/10.51847/Y4s7rY0dzz

Introduction

Molecular decision-making in drug discovery is distributed across several scientific questions that are often modeled separately. A candidate may be represented as a two-dimensional graph for property prediction, as a three-dimensional pose for target binding, as an ensemble for conformational analysis, as a substrate or inhibitor for metabolism, and as a collection of alerts or inferred interactions for toxicity and off-target assessment. Each representation captures only a conditional view of the molecule. The unresolved problem is therefore not merely how to predict more endpoints, but how to maintain a coherent, revisable, and evidence-linked account of what is currently believed about the molecule across these domains. Without such an account, predictions generated at different stages may be difficult to reconcile, update, reproduce, or restrict to appropriate decisions.

Corresponding Author: Laura Keller; Department of Digital Molecular Twins and Living Accounts, Faculty of Pharmacy, ETH Zurich, Zurich, Switzerland. E-mail: laura.keller@pharma.ethz.ch

Contemporary molecular machine learning is largely organized around task-specific datasets, representations, models, and metrics, while pharmaceutical development requires coordination among chemical, biological, experimental, and translational considerations [1-3]. A benchmark may establish comparative performance under a defined split, and a property model may estimate an endpoint from a particular molecular encoding, but neither result automatically explains how the estimate should interact with conformational evidence, target-state uncertainty, metabolism, safety liabilities, or new experimental findings. Optimizing one metric can therefore produce a technically improved model without resolving whether the corresponding molecular account is scientifically complete, transferable to a new chemical series, or useful for a particular project decision.

Digital-twin concepts provide a relevant but incomplete precedent. In medicine, digital twins have been framed as computational counterparts that integrate heterogeneous data and support conditional reasoning about an evolving biological system [4]. More recent discussion of digital twins in drug discovery and development similarly emphasizes opportunities to connect models, evidence, and decisions across changing development contexts [5]. However, transferring the term to molecular design without an explicit architecture risks relabeling an ordinary prediction model, simulation workflow, or data dashboard as a twin. A molecular twin must therefore be distinguished by the persistence of its state, the traceability of its evidence, the explicitness of its update rules, and the boundedness of its permitted uses.

This article develops the digital molecular twin as a proposed, versioned, molecule-specific computational state system. Its purpose is to maintain conditional accounts of molecular identity, structural ensembles, dynamics, target binding, metabolism, toxicity, and off-target liabilities; update those accounts when simulation or experimental evidence becomes available; support counterfactual lead-optimization queries; propagate uncertainty across connected components; and declare the decisions for which an output is or is not sufficiently supported. The contribution is architectural rather than empirical. It does not claim that a universal twin has been implemented or validated. Instead, it specifies the constructs, relationships, update logic, validation requirements, and decision-use boundaries needed to determine whether a future system merits the designation.

Why Static Molecular Representations Cannot Support Living Predictions

A static molecular representation is any encoding treated as a sufficiently complete input for a prediction without preserving the conditions, alternatives, and evidence history that determine its interpretation. Examples include a canonical string, a fixed graph, a descriptor vector, a single conformer, or a learned embedding. These encodings can be highly useful, but their adequacy depends on the task, data regime, and model with which they are paired. Comparative analyses of molecular representations show that performance varies materially with representation type, learning architecture, dataset scale, and endpoint, undermining the assumption that one encoding can serve as a universally sufficient molecular state [6]. The proposed twin therefore treats each representation as a view generated under declared assumptions rather than as the molecule itself.

Representation learning does not remove this conditionality. A learned embedding may compress regularities that are predictive within its training distribution, yet it may fail to preserve the chemical determinants required for transfer to a new scaffold, assay, target family, or physical regime. Critical evaluation of molecular representation learning has shown that strong task performance does not by itself demonstrate that an encoding contains transferable chemical knowledge [7]. In a living account, the embedding must consequently be accompanied by its training domain, intended endpoint, uncertainty behavior, version, and known failure modes. An embedding may contribute evidence to a twin state, but it cannot replace the state ontology, provenance record, or update mechanism.

Static prediction also conceals the interaction among representation, dataset design, pretraining, splitting strategy, and model architecture. Systematic comparison of these elements indicates that apparent gains cannot always be attributed to the molecular representation alone, because data composition and evaluation design may exert equally important effects [8]. A living prediction must therefore retain the conditions under which an estimate was produced and provide a mechanism for revising that estimate when the conditions change. A value generated from one assay definition, species, target construct, protonation assumption, or chemical domain should not be silently reused as though it were invariant across contexts.

The strongest challenge to a smooth and static molecular account is local discontinuity. Activity-cliff analyses demonstrate that small structural changes can be associated with large changes in biological activity and that molecular machine-learning models may fail precisely in these chemically consequential neighborhoods [9]. Such behavior matters because lead optimization proceeds through local modifications rather than through random sampling of chemical space. The proposed digital molecular twin must therefore preserve alternative states, local analog evidence, contradictory observations, and uncertainty around each proposed change. It must also distinguish a model's smooth interpolation from experimentally supported molecular behavior. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop why static molecular representations cannot support living predictions within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Why static molecular representations cannot support living predictions

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Molecular identity and form	Which molecular entity is actually being represented?	Tautomers, protonation states, stereoisomers, isotopic forms, salts,	Establishes the identity anchor for all later structural,	A canonical structure may be treated as a complete and	A registration string identifies a molecular form but does not establish which form

		and preparation conditions can alter the relevant state.	binding, metabolic, and safety claims.	condition-independent identity.	dominates in every experimental or biological environment.
Representation choice	Which properties of the molecule are retained or discarded by the encoding?	Molecular graphs, descriptors, conformers, and learned embeddings preserve different information and exhibit task-dependent performance [6].	Justifies maintaining several linked representations rather than selecting one universal encoding.	The highest-performing representation on one benchmark may be declared generally superior.	Representation suitability is conditional on endpoint, dataset, model, and context of use.
Learned molecular embedding	Does the learned representation transfer beyond its training regularities?	Predictive embeddings may capture dataset-specific correlations without preserving broadly transferable chemical determinants [7].	Requires embeddings to carry provenance, applicability, and uncertainty metadata.	Latent-space similarity may be interpreted as mechanistic or causal similarity.	An embedding is a model-dependent evidence object, not a complete or experimentally verified molecular state.
Data and evaluation context	Which aspects of performance arise from representation, data composition, splitting, or model architecture?	Molecular-property performance reflects interacting choices in representation, pretraining, dataset construction, and evaluation design [8].	Motivates storage of data lineage, task definition, split logic, and model version with every prediction.	Improvements caused by favorable evaluation design may be attributed to scientific understanding.	Benchmark performance is interpretable only within the conditions that generated it.
Local chemical discontinuity	How does the system behave when a small structural change produces a large endpoint change?	Activity cliffs expose failures of smooth similarity assumptions and model interpolation [9].	Grounds the need for local analog evidence and counterfactual uncertainty.	A visually minor modification may be assumed to cause a proportionally minor property change.	Structural proximity does not guarantee biological or pharmacological continuity.
Evidence updating	Can the representation change when new simulation or experimental evidence arrives?	Static encodings lack an inherent rule for reconciling new observations with prior estimates.	Introduces the need for a versioned evidence ledger and update operator.	New values may overwrite older values without preserving conflict, method, or source conditions.	Updating must create an auditable new state rather than erase the history of prior states.
Decision interpretation	For which scientific decision is a representation or prediction adequate?	Different decisions require different endpoints, error tolerances, comparators, and evidence standards.	Connects representation adequacy to later validation and decision-use boundaries.	A model valid for ranking may be used for mechanism, safety, or progression decisions.	Technical validity for one task does not confer general pharmaceutical usefulness.

State Variables across Structure, Dynamics, Binding, Metabolism, and Toxicity

The digital molecular twin is proposed as a system of connected state variables rather than a container of unrelated predictions. A state variable is a versioned, condition-dependent account of a scientifically meaningful aspect of the molecule, together with the evidence, assumptions, uncertainty, and provenance required to interpret it. Structural ensembles and molecular dynamics show why molecular state cannot be reduced to one optimized geometry; binding free-energy methods show that target interaction is conditional on thermodynamic paths, sampled configurations, and protocol assumptions; and predictive ADMET research demonstrates that disposition and safety require several connected endpoints rather than a single developability score [10-12]. The state variables are therefore not interchangeable outputs. Each represents a distinct evidentiary question and must retain its own validity conditions.

The structural state should begin with molecular identity but extend to distributions of conformers, protonation and tautomeric alternatives, environment-dependent geometries, and experimentally or computationally supported populations. The dynamical state should describe transitions among relevant conformational states, their approximate populations, and the conditions under which they were inferred. Molecular-dynamics simulation can reveal structural fluctuations and rare transitions that are absent from a static structure, but the resulting account remains conditional on force fields, simulation

setup, sampling sufficiency, and the appropriateness of the modeled environment [10]. Accordingly, a twin should not store "the conformation" as a privileged object. It should store a set of supported structural states, their relations, and explicit indications of what has not been adequately sampled.

The binding state should represent target context, target conformation, proposed pose, interaction pattern, thermodynamic estimate, kinetic evidence where available, and the experimental or computational conditions supporting each claim. Relative binding free-energy calculations can inform comparisons within related ligand series, but their interpretation depends on suitable poses, mappings, sampling, force fields, and protocol consistency [11]. Binding should therefore be recorded as a conditional relationship between a molecular state and a target state, not as a permanent scalar property of the ligand. The metabolism, disposition, and toxicity states should likewise remain endpoint-specific. Predictive ADMET frameworks can help organize physicochemical behavior, enzyme interactions, transport, metabolites, exposure-related liabilities, and safety-relevant off-targets, but computational coverage cannot be equated with experimental confirmation or human pharmacokinetic and toxicological validity [12].

The off-target state extends the molecular account beyond the intended target to alternative proteins, pathways, cellular responses, phenotypes, and higher-level biological signatures. The Chemical Checker illustrates how small molecules can be organized across chemical, target, network, cellular, and clinical signature spaces, thereby demonstrating that molecular similarity can be defined at several biological levels [13]. Such relationships are useful for generating selectivity and liability hypotheses, but they do not establish mechanism, causality, exposure relevance, or clinical consequence. In the proposed twin, off-target associations should therefore be linked to their source type, biological level, confidence, and experimental status. The resulting state system is not a single vector with more dimensions; it is a typed, evidence-linked network in which structure, dynamics, binding, metabolism, toxicity, and off-target context may constrain one another while retaining distinct uncertainties and validation requirements.

Architecture of the Digital Molecular Twin

The proposed architecture has six interacting layers: molecular identity, scientific state, evidence, updating, query, and governance. The identity layer distinguishes constitutional, stereochemical, protonation, tautomeric, isotopic, and preparation-specific forms. The scientific-state layer maintains separate but connected accounts of structure, dynamics, binding, metabolism, toxicity, and off-target risk. The evidence layer records the source, method, conditions, date, quality, and affected variables for every computational or experimental contribution. Biomedical knowledge-graph research shows why typed entities, relations, ontologies, and provenance are necessary when heterogeneous drug-discovery evidence is integrated [14]. The twin therefore requires more than a common storage format: it requires explicit semantics for what each node, relation, prediction, measurement, and uncertainty statement means.

A relational substrate connects the molecule to targets, protein states, pathways, enzymes, metabolites, assays, adverse outcomes, and biological contexts without collapsing these relationships into a single score. Integrated biomedical knowledge graphs demonstrate that drugs, genes, diseases, phenotypes, pathways, and other entities can be linked through typed relations while retaining their distinct identities [15]. In the proposed twin, however, an edge must also carry evidentiary status. A literature-derived association, a model prediction, a simulation-derived interaction, and an experimentally confirmed relationship cannot be represented as epistemically equivalent. Relation type, source conditions, confidence, contradiction status, and update history must remain visible.

The evidence ledger provides the architecture's persistence. New findings are appended as versioned evidence objects rather than used to overwrite previous values silently. Extensible knowledge-graph systems illustrate how experimental data, database records, literature evidence, and analytical results can be added while preserving their source context [16]. For the molecular twin, each update should create an auditable state transition that identifies what changed, why it changed, which evidence triggered the change, which downstream variables were recalculated, and which earlier claims remain unresolved. Conflicting results should coexist until a declared reconciliation rule or new experiment justifies a revised interpretation.

The query layer supports forward, inverse, and counterfactual reasoning. Forward queries ask how the current molecular state relates to a property or liability; inverse queries ask which structural characteristics may satisfy a desired set of conditions; counterfactual queries ask how a defined modification may alter connected states. Bidirectional molecular models show that structure-to-property and property-to-structure operations can be represented within one computational system [17]. Nevertheless, bidirectionality alone does not establish a digital twin. The defining contribution is the coupling of such queries to versioned evidence, state-specific uncertainty, update logic, and decision-use restrictions. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop architecture of the digital molecular twin within the article's central argument.

Table 2. Components, relationships, uncertainties, and validation needs within Architecture of the digital molecular twin

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Molecular identity layer	Registered structure, stereochemistry, protonation, tautomeric	Establishes the specific molecular entity to which	Versioned identity record with unresolved alternatives	Ambiguous or interconverting forms may be	Analytical and registration consistency under relevant conditions

	form, salt or preparation context	evidence is assigned		collapsed incorrectly	
Structural-state layer	Experimental structures, conformer generation, quantum calculations, molecular simulations	Represents plausible geometries and structural ensembles	Condition-linked conformer set and population hypotheses	Sampling, force-field, and environmental dependence	Agreement with orthogonal structural observables and reproducible sampling
Dynamical-state layer	Trajectories, kinetic models, time-resolved measurements	Represents transitions, metastable states, and timescales	State populations and transition relationships	State definitions and rare-event coverage may be incomplete	Convergence analysis and comparison with kinetic or spectroscopic evidence
Binding-state layer	Target structures, poses, interaction models, free-energy estimates, binding assays	Maintains target- and condition-specific interaction hypotheses	Binding-state record with pose, affinity, kinetics, and assumptions	Pose error, target-state uncertainty, protocol dependence	Orthogonal structural, thermodynamic, kinetic, and functional confirmation
Metabolism and disposition layer	Physicochemical data, enzyme and transporter evidence, metabolite predictions and assays	Organizes exposure-related and transformation liabilities	Endpoint-specific ADME and metabolite state	Species, assay, and exposure-context differences	Prospective endpoint-level calibration and experimental confirmation
Toxicity and off-target layer	Structural alerts, profiling data, biological signatures, pathway evidence, toxicology assays	Maintains safety and selectivity hypotheses	Evidence-linked liability map	Incomplete target coverage and uncertain causal relevance	Exposure-aware orthogonal testing and mechanistic follow-up
Evidence ledger	Simulation outputs, experimental observations, literature-derived relations, metadata	Preserves provenance, conditions, evidence type, and contradiction status	Immutable evidence objects linked to affected states	Missing metadata and incompatible protocols	Reproducible source tracing and completeness auditing
Update operator	Current twin version, new evidence, weighting and reconciliation rules	Produces an auditable revised state	New version with explicit state differences	Update rules may propagate bias or unjustified certainty	Reproducibility, sensitivity analysis, and prospective update evaluation
Query layer	Current state, user-defined constraints, structural or evidentiary intervention	Supports forward, inverse, and counterfactual reasoning	Conditional hypothesis with affected states and uncertainties	Out-of-domain queries may appear chemically plausible	Local validity checks, constraint tests, and experimental falsification
Uncertainty and governance layer	Model uncertainty, data variability, sampling limits, calibration history, intended use	Controls interpretation, escalation, and permitted decisions	Use-bounded output with uncertainty decomposition	Dependence among errors may be unknown	Decision-specific calibration and accountable human review

Figure 1 presents the living digital molecular twin, showing how the article's key components and boundaries are connected within architecture of the digital molecular twin.

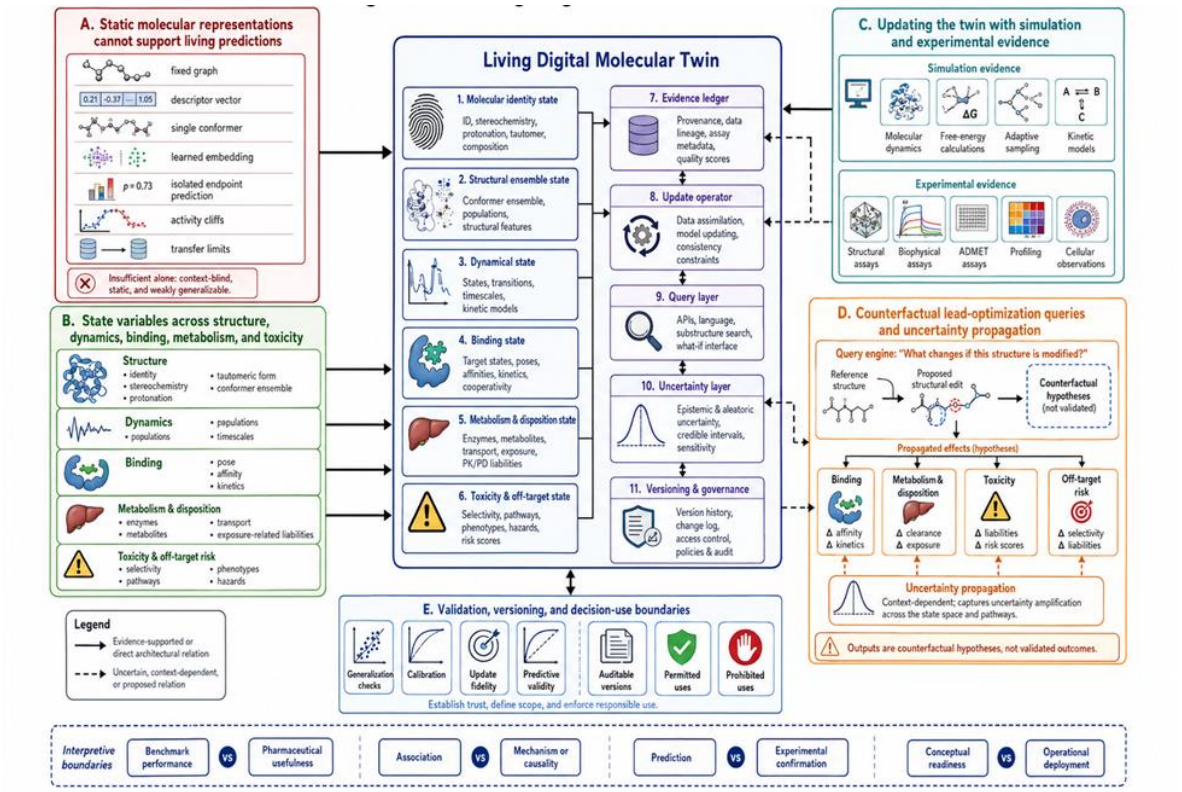

Figure 1. Living Digital Molecular Twin.

The figure is an original conceptual synthesis that organizes the article's central contribution across static molecular representations cannot support living predictions, state variables across structure, dynamics, binding, metabolism, and toxicity, state variables across structure, architecture of the digital molecular twin, updating the twin with simulation and experimental evidence, counterfactual lead-optimization queries and uncertainty propagation. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Updating the Twin with Simulation and Experimental Evidence

A molecular twin becomes living only when new evidence can modify its state through explicit and reproducible rules. Simulation contributes sampled conformations, transition hypotheses, interaction energies, and mechanistic possibilities, whereas experiments contribute observations under defined assay, structural, biophysical, cellular, or pharmacological conditions. Machine learning can accelerate molecular simulation and approximate otherwise expensive physical calculations, but the resulting outputs remain conditional on training coverage, model assumptions, and the physical fidelity of the learned representation [18]. Simulation-derived information should therefore enter the evidence ledger as model-conditioned evidence, not as an unquestioned observation.

Dynamic updating requires a state model capable of representing populations and transitions. Markov state models provide a formal example in which molecular trajectories are decomposed into metastable states and transition probabilities [19]. Their use also demonstrates that an apparently precise kinetic account depends on state construction, lag time, sampling, and validation choices. The twin should consequently preserve alternative state decompositions when they remain scientifically plausible, record which observables support each model, and avoid converting an under-sampled trajectory into a definitive description of molecular kinetics.

Evidence acquisition may also be adaptive. When particular conformations, binding events, metabolites, or off-target hypotheses dominate decision uncertainty, additional simulation or experimentation may be directed toward those gaps. Adaptive sampling can improve exploration of molecular processes, but studies of spontaneous ligand binding show that targeted sampling does not guarantee that all relevant events will be captured [20]. The twin's update mechanism should therefore distinguish reduced uncertainty from complete coverage. A newly sampled event may alter a state, yet the absence of an event in the current evidence cannot be treated as evidence that it is impossible.

Experimental observations can constrain, revise, or challenge simulation-derived states. Augmented Markov models demonstrate that molecular simulations and experimental measurements can be combined within a formal kinetic framework [21]. This supports a broader update principle: new evidence should affect only the variables and relationships for which a defensible connection exists. A binding assay may revise an interaction hypothesis without validating a pose; a metabolite-identification study may update a transformation pathway without establishing human exposure; and a cellular phenotype may

modify an off-target hypothesis without proving causality. Every update must preserve the distinction between what was observed, what was inferred, what was simulated, and what remains unresolved.

Counterfactual Lead-Optimization Queries and Uncertainty Propagation

The counterfactual query layer asks how a defined intervention could alter the connected molecular state. A query may substitute a functional group, constrain a conformation, alter protonation assumptions, change the target state, or introduce a new experimental observation. Its output should not be a single optimized score. It should be a structured account of which state variables are expected to change, which remain unsupported, how uncertainty propagates, and which evidence would be required to test the proposal. Research on uncertainty quantification in drug design shows that model, data, and applicability uncertainties must be distinguished, and comparative studies demonstrate that uncertainty estimators behave differently across molecular tasks and distribution shifts [22-24].

Molecular counterfactual methods provide a practical basis for generating nearby structures that alter a model prediction while retaining chemical constraints [25]. Within the twin, such structures are interpreted as hypotheses rather than candidate recommendations. A counterfactual that improves predicted affinity may worsen conformational accessibility, metabolic stability, transporter interaction, toxicity, or off-target risk. It may also fall outside the domain in which one or more component models are calibrated. The query engine must therefore return a multidomain consequence map rather than rank compounds solely by the endpoint that initiated the query.

Uncertainty propagation should be explicit at the level of each state and relation. Aleatoric variability, model uncertainty, sampling incompleteness, measurement error, dataset shift, and identity ambiguity should not be compressed into an undifferentiated confidence score. Dependencies also matter. A binding prediction and an off-target prediction may share training data or representation errors, while metabolic and toxicity estimates may depend on the same unverified exposure assumptions. The architecture should identify known dependencies, avoid assuming independence when evidence is absent, and show when an apparently narrow aggregate uncertainty is produced by unsupported combination rules.

The decision value of a counterfactual query lies in its capacity to expose trade-offs and specify discriminating evidence. A useful output may state that a proposed substitution is compatible with one binding hypothesis, increases uncertainty in a metabolic pathway, leaves a toxicity domain unevaluated, and should be tested through a defined assay or simulation. This is more scientifically defensible than presenting the modification as an optimized solution. Counterfactual reasoning in the proposed twin is therefore a mechanism for organizing falsifiable lead-optimization hypotheses, not for establishing causal effects, synthesizability, developability, safety, or therapeutic value.

Validation, Versioning, and Decision-Use Boundaries

Validation must begin by determining whether the evaluation design measures the intended form of generalization. Ligand-based benchmarks can reward memorization when closely related compounds appear across training and test partitions [26]. A molecular twin should therefore be evaluated through increasingly demanding conditions, including scaffold novelty, temporal separation, target-state change, assay transfer, chemical-series transfer, and sequential evidence updates. Retrospective performance on a favorable split may support a narrow technical claim, but it cannot establish prospective usefulness.

Versioning must make every output reproducible. Standardized modeling pipelines show how data preparation, splitting, model configuration, training, evaluation, and stored artifacts can be made traceable [27]. The twin extends this requirement from models to the whole molecular account. Every version should record identity assumptions, evidence objects, state estimates, model and software versions, calibration status, update rules, unresolved contradictions, and permitted uses. A later version should not erase the fact that an earlier decision was made from less complete or differently interpreted evidence.

Pharmaceutical usefulness must be distinguished from technical performance. Critical analyses of artificial intelligence in drug discovery show that benchmark success, retrospective prediction, and compelling demonstrations may fail to translate into medicinal-chemistry impact or project advancement [28]. The twin should therefore declare its decision-use boundary with every output. A state estimate may be suitable for generating an experiment, prioritizing an evidence gap, or comparing analogues, while remaining unsuitable for stopping a program, declaring a compound safe, inferring human exposure, or replacing accountable expert judgment.

Predictive validity is ultimately defined relative to a specific decision, comparator, downstream consequence, and operating context [29]. Validation of the digital molecular twin must therefore occur at several levels: identity fidelity, state accuracy, update reproducibility, uncertainty calibration, counterfactual consistency, generalization under novelty, and decision impact. Passing one level does not validate the others. **Table 3** organizes the evidence, constructs, and interpretation boundaries needed to develop validation, versioning, and decision-use boundaries within the article's central argument.

Table 3. Evaluation, implementation, and research priorities arising from The Digital Molecular Twin as a Living Account of Structure, Dynamics, Binding

Evaluation dimension	What must be tested	Suitable evidence or assessment	Failure signal	Interpretive limitation	Decision relevance
Molecular identity fidelity	Whether evidence is assigned to the correct	Analytical confirmation,	Predictions refer to a different form	Identity consistency does	Required before any downstream

	molecular form and preparation context	stereochemical audit, protonation and tautomer review	from the tested material	not establish environment-specific prevalence	molecular comparison
Structural-state adequacy	Whether relevant conformers and ensembles are represented	Replicate simulations, experimental structures, spectroscopic observables	Important states appear only after protocol or seed changes	Apparent convergence may remain condition-specific	Determines whether binding and reactivity hypotheses have a credible basis
Dynamical-state validity	Whether populations and transitions are reproducible and experimentally plausible	Kinetic modeling, held-out observables, time-resolved measurements	State populations or timescales change materially under reasonable analysis choices	Simulation timescales may not map directly to biological conditions	Governs use of kinetic and accessibility claims
Binding-state validity	Whether interaction claims survive orthogonal methods	Structural, thermodynamic, kinetic, and functional assays	Favorable prediction is unsupported by independent evidence	Binding confirmation does not establish functional efficacy	Supports target-engagement hypotheses, not therapeutic claims
ADME and toxicity validity	Whether endpoint estimates transfer to the intended assay, species, and exposure context	Prospective in vitro and in vivo assessment, endpoint-specific calibration	Missing or shifted liabilities emerge outside the modeled domain	Cross-species and exposure translation remain conditional	Supports liability triage but not safety clearance
Update fidelity	Whether new evidence produces a reproducible and scientifically justified state revision	Pre-update and post-update audits, alternative update-rule comparison	Unrelated variables change or prior evidence disappears without justification	Reproducible updating may still encode a flawed causal assumption	Determines whether the twin can be treated as a living account
Uncertainty calibration	Whether stated uncertainty corresponds to observed prediction error	Calibration curves, coverage assessment, temporal or prospective holdouts	High-confidence errors cluster in decision-relevant regions	Calibration can degrade after domain shift	Controls escalation, abstention, and evidence-generation decisions
Counterfactual validity	Whether proposed molecular changes yield coherent, testable, and locally supported consequences	Matched-series evaluation, synthesis review, orthogonal modeling, prospective assays	Suggested changes violate chemistry or fail across connected endpoints	Local plausibility does not establish causal or therapeutic benefit	Supports hypothesis generation, not autonomous lead selection
Generalization validity	Whether performance persists under realistic novelty	Scaffold, temporal, target, and project-level holdouts	Performance collapses as similarity to training data decreases	Retrospective tests remain imperfect proxies for future discovery	Sets the boundary for transferring outputs across projects
Decision impact	Whether use of the twin improves a specified decision relative to alternatives	Prospective human-in-the-loop studies, comparator workflows, error-cost analysis	More information produces no benefit or worsens decisions	Impact is local to the decision, team, and operating environment	Required before operational or portfolio-level use

Limitations and Boundary Conditions

The proposed digital molecular twin is a conceptual architecture, not an empirically validated computational platform. Its state ontology has not been shown to be complete, universally transferable, or equally appropriate across therapeutic modalities, target classes, chemical spaces, or stages of discovery. Some state variables may be well supported for one molecule and almost entirely unobserved for another. The architecture also depends on the quality of underlying models and experiments; organizing weak, biased, or incompatible evidence more carefully does not transform it into strong evidence.

The relationships among state variables may be correlational, mechanistic, causal, or merely operational, and the architecture must not blur these distinctions. A structural modification associated with a predicted toxicity change does not establish that the modification causes the biological outcome. Likewise, an experimentally observed endpoint does not necessarily identify the molecular mechanism producing it. Knowledge-graph connections, simulation trajectories, learned embeddings, and counterfactual explanations can help organize hypotheses, but each remains bounded by coverage, assumptions, and validation conditions [14]. The twin should therefore preserve epistemic labels rather than present every connected relation as a mechanistic pathway.

Implementation may create additional risks. A complex twin can appear authoritative because it aggregates many models, datasets, and visual relations, even when the components share errors or rely on the same uncertain assumptions. Automation may also encourage users to treat missing evidence as favorable evidence or to interpret a generated molecular modification as an endorsed design. These risks require abstention rules, visible uncertainty, contradiction handling, and accountable human oversight. The proposed construct cannot establish clinical benefit, regulatory acceptability, manufacturing feasibility, portfolio value, or deployment readiness, and it should not be used as an autonomous decision tool.

Research Agenda and Implementation Priorities

The first priority is to define a minimum molecular-twin ontology and reporting standard. Research should determine which identity, structural, dynamical, binding, metabolic, toxicity, and off-target variables are essential for different decision contexts; how missingness and contradiction should be encoded; and which provenance fields are necessary for reproducibility. Longitudinal benchmark designs should release evidence sequentially so that systems are evaluated not only on static prediction but also on state revision, calibration drift, contradiction recovery, and the preservation of prior versions. Such benchmarks should use scaffold, temporal, project, and assay shifts to reduce the risk of rewarding memorization [26].

The second priority is prospective validation of update and counterfactual operations. Comparative studies should test alternative rules for combining simulations with experimental observations, propagating dependent uncertainties, and selecting the next evidence-generating action. Counterfactual molecular modifications should be assessed in matched design–make–test series that record not only the initiating endpoint but also connected changes in binding, metabolism, toxicity, and off-target behavior. Failure cases and negative results should remain part of the evidence ledger. Without such testing, the architecture may produce internally coherent narratives that do not improve scientific decisions.

The third priority is staged implementation. An initial system should be limited to evidence organization, provenance, state comparison, and explicit uncertainty display. A second stage could support bounded counterfactual queries and evidence-gap prioritization under mandatory expert review. Only after reproducibility, calibration, generalization, and human decision impact have been demonstrated should more consequential uses be considered. Reporting should identify the exact twin version, component models, data lineage, state coverage, unresolved conflicts, uncertainty method, validation domain, permitted uses, and prohibited inferences. This staged approach recognizes that predictive validity must be demonstrated for the decision being supported rather than inferred from technical sophistication alone [29].

Conclusion

The digital molecular twin is proposed as a living, versioned, and evidence-linked account of a molecule rather than as another static representation or isolated predictor. Its central contribution is to connect molecular identity, structural ensembles, dynamics, target binding, metabolism, toxicity, and off-target risk through explicit state variables, provenance-preserving updates, counterfactual queries, uncertainty propagation, and decision-use boundaries. The architecture is intended to make assumptions, contradictions, evidence gaps, and limits visible while preventing benchmark performance, model confidence, or molecular plausibility from being mistaken for experimental confirmation or pharmaceutical value. Its usefulness remains conditional on the development of shared ontologies, rigorous update rules, prospective validation, calibrated uncertainty, reproducible versioning, and accountable human oversight. The construct should therefore be treated as a research framework for organizing and testing molecular reasoning, not as a validated predictive model, autonomous optimization system, regulatory standard, or deployment-ready decision tool.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. 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
2. 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
3. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
4. Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, Gustafsson M, et al. Digital twins to personalize medicine. Genome Med. 2020;12(1):4. doi:10.1186/s13073-019-0701-3

5. Venkatapurapu SP, Clegg L, Nowojewski A, Kimko H, Olabode D, Sawant-Basak A, et al. Digital twins for accelerating drug discovery and development: Opportunities and challenges. Drug Discov Today. 2026;31(2):104617. doi:10.1016/j.drudis.2026.104617
6. 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
7. Song B, Zhang J, Liu Y, Liu Y, Jiang J, Yuan S, et al. A systematic review of molecular representation learning foundation models. Brief Bioinform. 2026;27(1):bbaf703. doi:10.1093/bib/bbaf703
8. Fooladi H, Vu TNL, Mathea M, Kirchmair J. Evaluating machine learning models for molecular property prediction: Performance and robustness on out-of-distribution data. J Chem Inf Model. 2025;65(19):9871-91. doi:10.1021/acs.jcim.5c00475
9. 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
10. Hollingsworth SA, Dror RO. Molecular dynamics simulation for all. Neuron. 2018;99(6):1129-43. doi:10.1016/j.neuron.2018.08.011
11. Cournia Z, Allen B, Sherman W. Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. J Chem Inf Model. 2017;57(12):2911-37. doi:10.1021/acs.jcim.7b00564
12. Fraser JS, Edgar S, Handly LN, Kosuri S, Chodera JD, Murcko M, et al. Mapping the avoid-ome: A systematic open-science approach to predictive ADMET. Nat Commun. 2026;17(1):4644. doi:10.1038/s41467-026-73410-8
13. Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, et al. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. Nat Biotechnol. 2020;38(9):1087-96. doi:10.1038/s41587-020-0502-7
14. Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, et al. A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. Brief Bioinform. 2022;23(6):bbac404. doi:10.1093/bib/bbac404
15. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10(1):67. doi:10.1038/s41597-023-01960-3
16. Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, et al. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol. 2022;40(5):692-702. doi:10.1038/s41587-021-01145-6
17. Chang J, Ye JC. Bidirectional generation of structure and properties through a single molecular foundation model. Nat Commun. 2024;15(1):2323. doi:10.1038/s41467-024-46440-3
18. Noé F, Tkatchenko A, Müller KR, Clementi C. Machine learning for molecular simulation. Annu Rev Phys Chem. 2020;71:361-90. doi:10.1146/annurev-physchem-042018-052331
19. Husic BE, Pande VS. Markov state models: From an art to a science. J Am Chem Soc. 2018;140(7):2386-96. doi:10.1021/jacs.7b12191
20. Betz RM, Dror RO. How effectively can adaptive sampling methods capture spontaneous ligand binding? J Chem Theory Comput. 2019;15(3):2053-63. doi:10.1021/acs.jctc.8b00913
21. Olsson S, Wu H, Paul F, Clementi C, Noé F. Combining experimental and simulation data of molecular processes via augmented Markov models. Proc Natl Acad Sci U S A. 2017;114(31):8265-70. doi:10.1073/pnas.1704803114
22. 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
23. Parrondo-Pizarro R, Lanini J, Rodríguez-Pérez R. Uncertainty quantification in molecular machine learning for property predictions under data shifts. J Chem Inf Model. 2026;66(2):923-35. doi:10.1021/acs.jcim.5c02381
24. 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
25. 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
26. 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
27. Minnich AJ, McLoughlin K, Tse M, Deng J, Weber A, Murad N, et al. AMPL: A data-driven modeling pipeline for drug discovery. J Chem Inf Model. 2020;60(4):1955-68. doi:10.1021/acs.jcim.9b01053
28. 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
29. Scannell JW, Bosley J, Hickman JA, Dawson GR, Trübel H, Ferreira GS, et al. Predictive validity in drug discovery: What it is, why it matters and how to improve it. Nat Rev Drug Discov. 2022;21(12):915-31. doi:10.1038/s41573-022-00552-x