Pharmacophore

ISSN-2229-5402

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

EVERY MOLECULAR PREDICTION NEEDS AN EVIDENCE PASSPORT LINKING CONFIDENCE, APPLICABILITY, PROVENANCE, AND EXPERIMENTAL CONSEQUENCE

Hiroshi Yamamoto[1]*, Francesca Greco[2], Andrea Bianchi[3], Ryota Nakamura[1]

1. ***Department of Molecular Prediction and Evidence Passports, Graduate School of Pharmacy, Kyoto University, Kyoto, Japan.***
2. ***Department of Confidence and Applicability Assessment, Faculty of Pharmacy, University of Bologna, Bologna, Italy.***
3. ***Department of Provenance and Experimental Consequence Tracking, Faculty of Pharmaceutical Sciences, University of Milan, Milan, Italy.***
4. ***Kabardino-Balkarian State University named after H.M. Berbekov, Kabardino-Balkarian Republic, Russia.***
5. ***North Caucasus Federal University, Stavropol, Russia.***

ARTICLE INFO

Received:
14 September 2024
Received in revised form:
09 November 2024
Accepted:
11 November 2024
Available online:
28 December 2024

Keywords: Molecular prediction, Evidence provenance, Applicability domain, Uncertainty quantification, Data lineage, Experimental decision-making

ABSTRACT

Molecular predictions increasingly influence which compounds are screened, synthesized, optimized, or advanced, yet the evidential conditions surrounding those predictions are often separated from the outputs that enter experimental workflows. A probability, rank, or predicted property can therefore travel farther than its calibration evidence, applicability limits, data provenance, model version, and intended use. This article develops a proposed molecular prediction evidence passport as an original prediction-governance construct for keeping those elements attached to a bounded prediction event. The approach synthesizes methodological insights from molecular machine learning, uncertainty quantification, biomedical data stewardship, reporting guidance, and pharmaceutical decision practice. The passport is defined as a versioned, machine-readable and human-interpretable record that identifies the prediction object; preserves model, code, representation, and data lineage; distinguishes model confidence from calibrated uncertainty; records applicability conditions and unresolved uncertainty sources; states the experimental decision the prediction may inform; and links subsequent experimental consequences back to the originating record. The central contribution is not another performance metric, but an architecture for preventing benchmark performance, explanation, plausibility, or organizational approval from being misread as experimental confirmation, mechanism, therapeutic value, or deployment readiness. The construct may support more inspectable handoffs across screening, optimization, and candidate selection, while enabling later audit of what was known, assumed, changed, and acted upon. Its value remains conditional on schema design, semantic interoperability, data quality, human oversight, and prospective evaluation. The passport is therefore presented as a testable governance proposal rather than a validated standard, regulatory artifact, or autonomous decision system.

To Cite This Article: Yamamoto H, Greco F, Bianchi A, Nakamura R. Every Molecular Prediction Needs an Evidence Passport Linking Confidence, Applicability, Provenance, and Experimental Consequence. Pharmacophore. 2024;15(6):66-75. https://doi.org/10.51847/WvgrULwuFS

Introduction

Machine learning is increasingly embedded in drug discovery, but meaningful impact remains conditional on well-defined questions, suitable evidence, realistic workflow integration, and downstream decision quality [1–3]. This condition is especially important in computational pharmaceutical science because a molecular prediction rarely exists as an isolated scientific statement. It can redirect screening resources, determine which analogues are synthesized, prioritize assays, influence target or candidate narratives, and shape later interpretations of failure or success. Once a prediction enters such a workflow, its practical meaning depends on more than the numerical output. The relevant model version, training evidence, molecular

Corresponding Author: Hiroshi Yamamoto; Department of Molecular Prediction and Evidence Passports, Graduate School of Pharmacy, Kyoto University, Kyoto, Japan. E-mail: yamamoto.hiroshi@kyoto-u.ac.jp.

representation, evaluation design, chemical neighborhood, uncertainty estimate, decision threshold, and proposed experimental response all contribute to whether the output can be used responsibly. When those elements are distributed across notebooks, repositories, reports, database records, and tacit team knowledge, the prediction becomes easier to reuse than to appraise. The resulting problem is not simply insufficient transparency. It is a loss of accountability between the conditions under which a prediction was generated and the experimental consequence later attributed to it.

This accountability problem begins with the data from which molecular models learn. Chemical and biological data are context-dependent, and availability alone does not establish suitability for a prediction task [4]. Bioactivity values can reflect different assay formats, target constructs, detection technologies, endpoint definitions, curation choices, salt and stereochemical conventions, replicate policies, and temporal versions of an underlying record. A model may therefore reproduce a regularity that is meaningful within one assembled dataset but misleading when transferred to another assay, chemical series, or decision stage. Conventional reporting often compresses this dependence into a dataset name and a summary metric, leaving downstream users unable to reconstruct what evidence the model encountered or which transformations changed its meaning. The distinction between data availability and data suitability is consequently central to prediction governance. A scientifically useful record must preserve not only where data originated, but also how they were selected, standardized, combined, split, and interpreted for the specific prediction object now being considered.

Interpretability does not close this gap by itself. Explainable artificial intelligence can expose associations, feature contributions, substructures, or learned response patterns, but an explanation does not independently establish a molecular mechanism or experimental validity [5]. An attribution map may be informative about model behavior while remaining unstable, representation-dependent, non-causal, or outside the region in which the model was adequately evaluated. Similar distinctions apply to predicted binding, activity, toxicity, permeability, or synthesizability: a plausible molecular rationale is not equivalent to measured target engagement, a causal pathway, developability, human safety, or therapeutic benefit. Prediction governance must therefore protect categorical boundaries among association, prediction, explanation, mechanism, experimental confirmation, clinical utility, and regulatory acceptability. Without those boundaries, additional model outputs can create an appearance of evidential depth while leaving the underlying decision unsupported.

This article proposes the molecular prediction evidence passport as an original conceptual response to that unresolved problem. The passport is not conceived as a universal score, static model card, database accession, or automated approval mechanism. It is a versioned evidence object attached to a bounded prediction event and designed to remain with that event as it moves through screening, optimization, and candidate-selection decisions. Its proposed functions are to identify what was predicted; preserve the model, code, representation, and data lineage; separate confidence from calibrated uncertainty; state the relevant applicability conditions; record unresolved uncertainty sources; specify the experimental decision the output may inform; and capture the consequence of acting, abstaining, or acquiring further evidence. By organizing these relationships, the article aims to make molecular predictions more inspectable without claiming that documentation can substitute for validation. The contribution is therefore a prediction-governance model whose components, evaluation needs, implementation priorities, and boundary conditions can be tested in future pharmaceutical workflows.

The Accountability Gap between a Prediction and Its Experimental Use

The accountability gap is the interval between a model output and the scientific responsibility attached to its experimental use. It appears whenever an output is transferred without the evidence needed to determine what the model learned, where it was evaluated, what uncertainty remains, and which action is justified. Benchmark performance can reflect dataset structure, split design, and representation choices rather than transferable molecular generalization [6–8]. This does not make benchmarking uninformative; it limits the claim that can be carried from a benchmark into a pharmaceutical decision. A score produced under a random split, for example, may answer whether a model interpolates within a familiar dataset more readily than whether it will support a new chemical series, assay configuration, target context, or temporal batch. Accountability therefore requires a record of the task definition, endpoint, metric, split rationale, comparator, preprocessing, and validation population before performance can be interpreted as evidence for a downstream use.

The gap is also local rather than merely global. Strong average performance can coexist with large errors for chemically similar compounds separated by steep activity changes [9]. Such activity cliffs demonstrate why a molecule may be nominally “in domain” under a broad similarity or leverage rule while remaining difficult to predict in its immediate neighborhood. A single global applicability flag is therefore too coarse for the proposed governance model. The relevant evidence includes neighborhood density, scaffold and series relationships, proximity to known cliffs, assay consistency, target coverage, and the behavior of comparable molecules under the same evaluation conditions. Even these indicators cannot prove that an individual prediction is correct. Their role is to qualify whether the model is interpolating within supported evidence, extrapolating across a recognized boundary, or operating in a region where uncertainty is not adequately characterized. The accountability gap widens when these distinctions are absent because downstream users may interpret a high score as stronger evidence than the local chemical landscape supports.

A third source of accountability loss is the substitution of computational achievement for experimental consequence. Generated or ranked molecules must be judged by experimental and medicinal-chemistry consequences, not novelty or computational score alone [10]. A proposed compound may be structurally unusual yet synthetically impractical, assay-interfering, unstable, nonselective, poorly exposed, or unsuitable for the project’s therapeutic hypothesis. Conversely, a modestly ranked molecule may be valuable because it tests a decisive mechanistic or structure–activity question. The relevant governance issue is

therefore not whether a model produced an impressive candidate, but what decision the prediction was intended to change and what evidence should follow from that change. A prediction used to reduce a docking library, nominate a compound for purchase, propose an analogue for synthesis, or support candidate selection carries different costs of error, evidence requirements, and oversight needs. Experimental consequence must be recorded as a bounded decision relation rather than inferred retrospectively from whether a project later succeeded.

Within the proposed model, the accountability gap is represented by six linked questions: what exactly was predicted; from which model and evidence state; for which molecule, endpoint, assay, and decision context; with what calibration and applicability support; under which unresolved uncertainties; and with what permitted experimental consequence. These questions convert accountability from a general aspiration into an inspectable relationship between evidence and action. They also expose overclaiming risks: benchmark performance may be mistaken for prospective usefulness, similarity for applicability, confidence for calibrated uncertainty, explanation for mechanism, prioritization for validation, and internal review for scientific or regulatory acceptance. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop the accountability gap between a prediction and its experimental use within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for the accountability gap between a prediction and its experimental use

Construct or Evidence Domain	Core Question	Scientific Basis	Role in the Article	Failure or Overclaiming Risk	Boundary Statement
Prediction Object and Task Definition	What exactly was predicted, for which molecule, endpoint, unit, and task?	Molecular outputs are interpretable only relative to a defined target variable, representation, and prediction context.	Establishes the smallest accountable unit to which the passport is attached.	A score is reused for a different endpoint, assay, molecule state, or decision.	Identifying a prediction does not validate its accuracy or pharmaceutical relevance.
Benchmark and Evaluation Context	Under which split, comparator, metric, and validation population was performance observed?	Benchmark outcomes depend on dataset structure, split logic, representation, and task design.	Defines the evidence that supports, or fails to support, transfer beyond the evaluation setting.	Benchmark performance is interpreted as prospective experimental usefulness.	Evaluation evidence is conditional on the population and design in which it was produced.
Data Suitability and Lineage	Which source records, assays, transformations, exclusions, and curation rules shaped the model?	Chemical and biological records are affected by assay context, standardization, duplication, missingness, and temporal change.	Connects the output to the evidence from which the model learned.	A named dataset is treated as sufficiently described or inherently suitable.	Data availability, accessibility, or scale does not establish suitability for a specific prediction task.
Molecular Representation Provenance	How were molecular identity, stereochemistry, protonation, conformation, and features encoded?	Representation choices can alter the information available to the model and the errors it produces.	Makes representation-dependent claims and reproduction requirements inspectable.	Equivalent molecules are encoded inconsistently or important molecular distinctions are lost.	Representation fidelity does not establish mechanistic fidelity.
Confidence, Calibration, and Applicability	What does the reported confidence mean, and is the molecule supported by evaluated chemical and biological conditions?	Point estimates, probabilities, calibration evidence, local neighborhoods, and distribution-shift indicators answer different questions.	Prevents confidence from functioning as an unqualified reliability claim.	High confidence is treated as calibrated certainty or proof of in-domain correctness.	In-domain status does not prove correctness, and out-of-domain status does not prove failure.
Interpretation And Mechanistic Status	Is the output an association, prediction, explanation, mechanistic hypothesis, or experimentally supported claim?	Model explanations describe model behavior and may not identify causal molecular mechanisms.	Protects categorical boundaries among computational interpretation and scientific evidence.	Feature attribution or plausibility is presented as causal or experimentally confirmed.	Explanation cannot substitute for mechanistic or experimental validation.
Intended Experimental Consequence	Which action may the prediction inform, and what evidence threshold is appropriate to that action?	Screening, synthesis, assay selection, optimization, and candidate nomination impose different costs of error and evidentiary requirements.	Links the prediction to a bounded decision rather than an unrestricted claim.	A low-consequence prioritization output is reused for a high-consequence decision.	The passport may inform a decision but does not make or authorize the decision autonomously.
Experimental Feedback and Stewardship	What happened after the prediction was used, and who reviewed or amended the record?	Prediction value can be evaluated only when subsequent experimental evidence and version changes remain traceable.	Closes the evidence loop and supports later reconstruction of actions and assumptions.	Only successful outcomes are retained, or later evidence is detached from the originating prediction.	Experimental confirmation remains assay- and context-specific and does not establish universal model validity.

Defining the Molecular Prediction Evidence Passport

The molecular prediction evidence passport is proposed as a unique, versioned record for one bounded prediction event or an explicitly defined prediction batch. Its scope begins with an identifiable output and ends with the experimental or review

consequence that the output is permitted to inform. Responsible prediction use requires lifecycle documentation of intended use, validation context, limitations, and monitoring [11]. Translated into molecular modeling, this means that a passport should identify the molecule or candidate set, endpoint, units, task, timestamp, model and code versions, representation and preprocessing, training and evaluation data lineage, performance and calibration evidence, applicability status, uncertainty sources, interpretive claims, decision owner, and planned experimental response. Mandatory fields should be minimal enough to complete routinely yet sufficient to prevent the prediction from becoming detached from its evidential conditions. Domain-specific extensions may then capture three-dimensional structure preparation, assay technology, target construct, molecular-dynamics protocol, generative objective, synthesis route, or other information required by the task.

The passport is an evidence container rather than an evidence score. A minimum-information structure can make model development, data, evaluation, and interpretation more inspectable [12]. It should therefore distinguish raw prediction from post-processed output, probability from calibrated probability, model explanation from supporting experimental evidence, and current passport state from superseded versions. Structured questions across the model lifecycle can additionally expose gaps in transparency, replicability, and demonstrated effectiveness [13]. In the proposed construct, an unanswered mandatory question is not silently converted into a neutral value; it is recorded as missing, unresolved, inaccessible, or not clearly reported. This design makes absence visible and allows decision rules to respond differently to unavailable provenance, untested calibration, ambiguous assay semantics, or an applicability warning. The passport should consequently preserve both supporting and conflicting evidence rather than compressing them into an undifferentiated confidence badge.

Machine readability is necessary because the record must travel across modeling environments, compound registries, assay systems, project repositories, and governance logs. Biopharmaceutical data reuse requires findability, accessibility, interoperability, and reusability supported by practical metadata [14]. For the passport, those principles imply persistent identifiers, controlled field definitions, explicit units and endpoint semantics, resolvable source links, recorded transformations, version relationships, and access conditions. They do not imply that every underlying record must be public or that FAIR data are automatically suitable for modeling. Rather, they establish the infrastructural conditions under which an authorized reviewer can understand where evidence came from, how it changed, and which version informed an action. **Figure 1** presents the molecular prediction evidence passport, showing how the article's key components and boundaries are connected within defining the molecular prediction evidence passport.

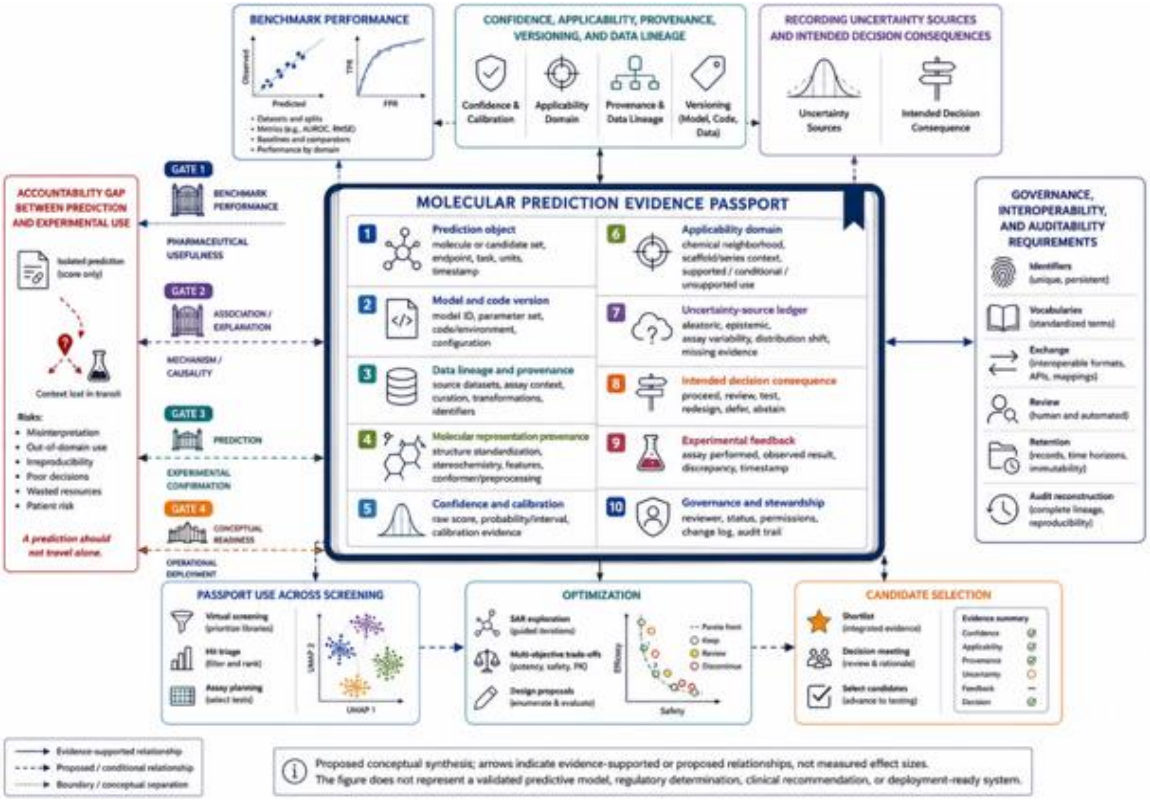

Figure 1. Molecular Prediction Evidence Passport.

The figure is an original conceptual synthesis that organizes the article's central contribution across accountability gap between a prediction and its experimental use, defining the molecular prediction evidence passport, confidence, applicability, provenance, versioning, and data lineage, data lineage, recording uncertainty sources and intended decision consequences, passport use across screening, optimization, and candidate selection. 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.

Interoperability also requires community-maintained standards rather than project-specific labels that only local users can interpret. Community-linked standards, repositories, and policies can make data and metadata practices discoverable and comparable [15]. A passport should therefore use registered vocabularies and identifier systems where suitable, while preserving extensions for task-specific evidence that existing standards do not capture. Compatibility must be evaluated semantically, not merely syntactically: two records may share a field named "activity" while referring to different assay endpoints, organisms, target constructs, time points, or normalization procedures. Governance fields should identify who created, reviewed, amended, or retired the record and which use was authorized, but such status must not be read as empirical validation or regulatory acceptance. The proposed passport is thus best understood as an accountable evidence interface: it can organize what is known, uncertain, missing, versioned, and intended, while leaving scientific judgment and experimental confirmation with qualified human teams.

Confidence, Applicability, Provenance, Versioning, and Data Lineage

Confidence should be recorded as a property of a specific model output rather than as a general declaration that the prediction is trustworthy. Comparative molecular studies show that uncertainty-estimation methods differ in their assumptions, calibration behavior, scalability, and response to distribution shift; no single method can therefore be treated as universally sufficient [16–18]. The proposed passport separates the point prediction from any probability, interval, ensemble dispersion, evidential parameter, or other confidence-related quantity attached to it. It must also identify how the quantity was produced, against which data it was evaluated, and whether calibration was examined globally, locally, or under a relevant chemical or temporal shift. This separation prevents a high numerical score from being interpreted automatically as calibrated uncertainty, in-domain reliability, or experimental certainty.

Applicability provides the contextual boundary within which confidence and calibration evidence may be interpreted. Uncertainty in drug design can arise from the observations, assay process, molecular representation, model, parameter estimates, chemical domain, or mismatch between the modeled endpoint and the intended decision [19]. The passport should therefore record applicability across more than chemical similarity. Relevant dimensions may include scaffold and series relationships, target family, assay technology, endpoint definition, molecular state, representation coverage, label density, temporal origin, and the presence of local activity discontinuities. Applicability status should be expressed as supported, conditionally supported, unsupported, or unresolved for the stated use, accompanied by the method and evidence used to reach that status. It should not be reduced to a universal binary boundary.

Provenance and versioning establish which evidence state produced the prediction. The proposed record should contain persistent or locally stable identifiers for the model, parameter set, source code, software environment, molecular input, featurization procedure, calibration model, decision threshold, and passport schema. Each substantive modification should produce a new linked version rather than silently overwriting the earlier record. Version relationships should identify whether the change affected data, code, model parameters, endpoint semantics, preprocessing, calibration, or governance status. A prediction reproduced from a later model version is not necessarily the same scientific object as the original prediction, even when the molecule and numerical result appear unchanged. Versioning therefore supports reconstruction of what was known and used at the time of a decision.

Data lineage extends provenance from named sources to the transformations that created model-ready evidence. A lineage record should trace source acquisition, inclusion and exclusion criteria, molecular standardization, duplicate handling, stereochemical treatment, assay harmonization, unit conversion, aggregation, imputation, censoring, label construction, split assignment, and subsequent corrections. It should also retain links to superseded or conflicting records when those records materially affect interpretation. The proposed passport does not assume that every lineage element will be publicly accessible; proprietary evidence may require controlled access. It does require that authorized reviewers can distinguish unavailable data from absent documentation and understand how the current prediction depends on earlier evidence states. Data lineage is thus not a decorative provenance statement but a scientific dependency map.

Recording Uncertainty Sources and Intended Decision Consequences

Point predictions alone do not communicate the reliability needed when a model output influences a consequential decision [20]. The proposed passport therefore contains an uncertainty-source ledger rather than a single undifferentiated uncertainty field. The ledger may identify observation noise, assay variability, label ambiguity, sparse chemical neighborhoods, model disagreement, parameter uncertainty, representation limitations, missing biological context, distribution shift, and unresolved conflicts among data sources. Each entry should state the evidence supporting its inclusion, the method used to characterize it, and whether it is quantified, qualitatively flagged, or not clearly estimable. Recording "unknown" is scientifically preferable to presenting an unsupported interval or interpreting model dispersion as a complete account of uncertainty.

Uncertainty becomes decision-relevant only when it is linked to an appropriate response. Communication of uncertainty can support review, abstention, escalation, second opinions, or the acquisition of additional evidence [21]. For molecular prediction, the passport should encode a bounded consequence such as proceed to routine screening, request expert review, conduct an orthogonal assay, obtain additional analogues, repeat model evaluation, redesign the candidate, defer the decision, or abstain from prediction-based prioritization. These consequence categories should not function as automatic commands. They should record the response the uncertainty evidence was intended to inform, the responsible role, the threshold or rationale applied, and whether alternative actions were considered.

Interpretability and uncertainty must also remain separate. Post hoc explanation should not be treated as a substitute for model transparency or as proof that a prediction is safe to act upon [22]. A substructure attribution, attention pattern, counterfactual molecule, or feature-importance profile may help a reviewer interrogate model behavior, but it may also be unstable, non-causal, or conditioned on the representation and explanation algorithm. The passport should identify the explanation method, target output, model version, stability evidence, and intended interpretive claim. It should explicitly label whether the explanation concerns model behavior, a structure–property hypothesis, or independently supported mechanism. This prevents explanatory plausibility from silently raising the evidential status of the prediction.

Trust in drug-discovery predictions requires explicit uncertainty characterization and cannot be inferred from point accuracy alone [23]. The proposed consequence record therefore links four elements: the uncertainty source, the applicability context, the cost of error, and the planned evidence response. A false negative during broad library triage, for example, has a different consequence from an unsupported safety reassurance during candidate selection. The passport should make that asymmetry visible without converting it into a universal numerical utility function. Decision consequences are project- and stage-dependent, and their legitimacy depends on scientific judgment, resource constraints, alternative evidence, and the reversibility of the action. The proposed structure organizes those dependencies; it does not establish the correct decision in advance.

Passport Use Across Screening, Optimization, and Candidate Selection

In early screening, molecular predictions frequently determine which compounds receive more expensive computation or experimental attention. Deep-learning-assisted docking illustrates how predictive triage can reduce the number of compounds subjected to full docking while retaining a selected subset for subsequent evaluation [24]. A screening passport should therefore record the candidate universe, library version, inclusion filters, model and docking protocols, acquisition or ranking rule, threshold, applicability checks, and the compounds excluded by the decision. Its experimental consequence is not a declaration of binding or activity. It is a documented resource-allocation decision that should remain reviewable when the screening strategy, chemical space, or model is later revised.

During molecular design and optimization, the passport should connect computational proposals to the assumptions that justified synthesis or testing. AI-guided identification of DDR1 inhibitors demonstrates the importance of linking generated or prioritized molecules to synthesis and experimental follow-up [25]. For each proposed analogue, the record should distinguish generated structure, predicted properties, design objective, relevant comparators, local chemical neighborhood, synthesizability assessment, and the experiment intended to test the design hypothesis. A favorable prediction may justify synthesis under one project constraint while remaining insufficient under another. The passport thus preserves why a compound was selected without implying that generation, novelty, or successful synthesis establishes therapeutic value.

Prospective experimental follow-up further demonstrates that different evidence layers answer different questions. Deep-learning-supported antibiotic discovery linked model ranking to laboratory testing of selected compounds, producing a sequence of evidence rather than a single validated prediction [26]. Within the proposed passport, a computational rank, primary assay result, orthogonal confirmation, mechanistic study, spectrum assessment, and in vivo observation would remain distinct but connected records. A positive result should not erase the uncertainties present when the selection was made, and a negative result should not be detached from the model version and decision rule that produced it. Recording both outcomes is essential for evaluating selection bias, learning from failure, and preventing retrospective inflation of model performance.

In active-learning and optimization cycles, each prediction can influence the next evidence acquired. Pool-based active learning for molecular screening demonstrates how iterative acquisition policies select compounds for further evaluation and thereby reshape the data available to later models [27]. The passport should identify the acquisition function, candidate pool, model version, uncertainty method, feasibility filters, selected experiment, result, and subsequent update. Candidate selection requires an even broader record because efficacy, selectivity, pharmacokinetics, formulation, safety, manufacturability, and strategic considerations cannot be collapsed into one model score. At that stage, passports from multiple endpoints may support a cross-functional decision, but they must preserve conflicting evidence and non-comparable uncertainties. The construct is intended to expose the basis and limits of nomination, not to automate candidate selection.

Governance, Interoperability, and Auditability Requirements

A passport becomes governable when its components, relationships, and statuses can be inspected independently of the interface that created them. Biomedical knowledge graphs provide a useful representational precedent because they can connect heterogeneous entities, relationships, and evidence without collapsing them into a single score [28]. A graph-compatible passport could link a prediction event to molecules, targets, assays, datasets, model artifacts, publications, decisions, experiments, and later versions through typed relations. Such representation would allow supporting, conflicting, derived, and superseded evidence to remain distinguishable. The knowledge graph itself would not validate the linked claims; its role would be to preserve their identities and relationships for authorized interpretation.

Interoperability also depends on structured molecular and assay records. Developments in ChEMBL toward direct deposition illustrate the importance of curated compounds, targets, assays, activities, identifiers, and links to source evidence [29]. Passport exchange should therefore use stable compound and target identifiers where possible, explicit endpoint and unit definitions, machine-readable assay descriptors, and provenance links to the underlying records. When an identifier is unavailable or contested, the record should preserve the submitted representation and document the mapping decision.

Syntactic compatibility is insufficient if two systems interpret activity, inhibition, exposure, or molecular identity differently. Interoperability must preserve scientific meaning, not merely transfer fields.

Integrated evidence platforms further show why prioritization outputs must retain access to their underlying evidence and release state. The Open Targets Platform integrates and normalizes multiple evidence sources for systematic target identification and prioritization [30]. A molecular prediction passport should similarly expose the evidence objects contributing to a derived assessment, the transformation or weighting logic, and the platform or model release used. Composite scores may be useful for navigation, but they should not erase whether the underlying evidence is genetic, experimental, computational, literature-derived, or conflicting. Governance requires users to be able to move from a summary output back to the evidence on which it depends.

Drug-discovery datasets differ in scope, identifiers, semantics, accessibility, and suitability for knowledge-graph integration, creating persistent interoperability and lineage challenges [31]. The passport therefore requires controlled vocabularies, identifier crosswalks, schema versions, extension rules, access controls, stewardship roles, change logs, retention policies, and procedures for correcting or retiring records. Auditability should permit reconstruction of who generated the prediction, which evidence and software versions were used, what decision was proposed, who reviewed it, what changed, and what experimental consequence followed. An audit trail documents process and responsibility; it does not certify that the scientific decision was correct. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop governance, interoperability, and auditability requirements within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Every Molecular Prediction Needs an Evidence Passport Linking Confidence, Applicability, Provenance, and

Component Or Layer	Inputs	Core Function	Expected Output	Uncertainty or Limitation	Validation Requirement
Passport Identity And Scope	Prediction identifier, molecule or batch identifier, endpoint, task, timestamp	Defines the bounded prediction event governed by the record	Unique, versioned passport object	Ambiguous batch boundaries or identifier collisions	Test whether independent users identify the same prediction object from the record
Model And Computational Versioning	Model identifier, parameters, code commit, environment, configuration	Preserves the computational state that produced the output	Reconstructable model-run specification	Nondeterminism, unavailable dependencies, undocumented modifications	Independent reproduction within a predefined numerical tolerance
Data And Representation Lineage	Source records, assay metadata, curation rules, molecular standardization, features	Traces how raw evidence became model-ready input	Traversable lineage from prediction to source evidence	Incomplete source metadata, inaccessible proprietary records, semantic inconsistency	Cross-system lineage-reconstruction and semantic-loss testing
Confidence And Calibration Record	Point prediction, probability or interval, calibration data, evaluation design	Distinguishes model output from calibrated reliability evidence	Context-qualified confidence statement	Calibration drift, local failure, distribution shift	External, temporal, scaffold-aware, and neighborhood-level calibration assessment
Applicability And Uncertainty Ledger	Domain tests, chemical neighborhood, model disagreement, assay variability, missing evidence	Identifies supported use conditions and unresolved uncertainty sources	Applicability status with typed uncertainty entries	Uncertainty sources may be entangled or incompletely identifiable	Controlled perturbation, out-of-domain, activity-cliff, and prospective-series evaluation
Intended Decision Consequence	Decision stage, permitted action, threshold, cost of error, alternatives	Restricts the prediction to a stated experimental or review use	Proceed, review, test, redesign, defer, or abstain rationale	Decision costs are project-specific and can change over time	Prospective comparison of decisions with and without passport information
Experimental Feedback	Planned experiment, protocol, result, discrepancy, timestamp	Connects prediction use to subsequent evidence acquisition	Versioned prediction–experiment relationship	Negative or failed experiments may be selectively omitted	Completeness audits and prospective outcome-capture studies
Interoperability Layer	Controlled vocabularies, persistent identifiers, ontology mappings, schema profile	Enables exchange without losing endpoint, assay, or evidence meaning	Machine-readable record with explicit mappings and warnings	Syntactic exchange may conceal semantic incompatibility	Round-trip exchange and semantic-equivalence testing across platforms
Governance And Audit Layer	Stewardship roles, review status, permissions, signatures, change log	Makes responsibility, authorization, amendment, and retirement visible	Reconstructable decision and change history	Organizational approval may be mistaken for scientific validation	Independent audit exercises and governance-failure simulations
Human Oversight Interface	Role-specific views, evidence summaries,	Supports expert review without erasing complexity	Documented human interpretation and decision rationale	Information overload or reductive traffic-light use	Human-factors studies measuring comprehension, error

unresolved warnings, access to source details	detection, burden, and disagreement

Limitations and Boundary Conditions

The proposed molecular prediction evidence passport is a conceptual governance model and has not been empirically validated as a complete or optimal schema. Its components arise from a synthesis of molecular modeling, uncertainty, reporting, provenance, and data-integration problems, but that synthesis does not demonstrate improved discovery outcomes. The construct may increase documentation without improving decision quality, or it may transfer existing ambiguity into structured fields. Uncertainty methods remain task-dependent, and even carefully evaluated approaches can fail under shifts or local chemical conditions not represented in their validation evidence [16]. The passport can expose such limitations, but it cannot remove them.

The model is also context-dependent. Screening, generative design, quantitative structure–activity modeling, docking, pharmacokinetic prediction, toxicology, formulation, and candidate selection require different metadata and evidentiary thresholds. A universal core schema may therefore need domain-specific modules, and excessive standardization could erase distinctions that the passport is intended to preserve. Dataset heterogeneity and incompatible semantics may prevent complete lineage integration even when technical exchange is possible [31]. Proprietary restrictions, legacy systems, missing source records, changing identifiers, and uncertain assay mappings may produce unavoidable gaps. Those gaps should be visible, but visibility alone does not make the remaining evidence sufficient.

Several forms of misuse remain possible. A completed passport could be treated as a certification badge, organizational approval could be mistaken for scientific validation, or a machine-readable consequence field could be converted into automated decision authority. Detailed provenance may also create a false impression that traceability guarantees data quality, causality, or reproducibility. The proposed model cannot establish molecular mechanism, synthesizability, developability, human safety, clinical effectiveness, regulatory acceptability, or project success. It cannot determine the appropriate level of risk for every organization, and it does not replace medicinal chemistry, assay science, pharmacology, toxicology, data stewardship, or governance expertise. Its proper boundary is narrower: preserving the evidential conditions and consequences of a prediction so that qualified users can appraise them.

Research Agenda and Implementation Priorities

The first research priority is to determine whether a stable minimum passport schema can be defined without oversimplifying heterogeneous molecular-prediction tasks. This requires structured consensus work involving computational scientists, medicinal chemists, assay specialists, pharmacologists, data stewards, software engineers, and governance experts. Candidate fields should be tested through realistic prediction cases in which participants identify missing evidence, interpret uncertainty, and decide whether the prediction can support a specified action. Evaluation should examine field completeness, interpretive consistency, detected overclaiming, review burden, and the frequency with which users appropriately request further evidence or abstain. The objective would be to identify a bounded core and extensible modules, not to impose one uniform record on every pharmaceutical workflow.

A second priority is prospective evaluation of passport utility. Studies should compare otherwise similar workflows with and without a structured passport while avoiding claims that documentation alone causes better discovery outcomes. Relevant assessments may include the ability to reproduce a prediction, identify the correct model and data version, detect out-of-domain use, recognize unsupported mechanistic language, reconstruct a selection decision, and connect experimental results to the originating prediction. Active-learning and iterative optimization settings are especially important because each selection changes the evidence available to later models. Studies should also retain negative, inconclusive, and failed experiments so that passport performance is not assessed only on successful examples.

Implementation should proceed in stages. An initial phase could use human-readable records attached to selected high-consequence prediction events. A second phase could add persistent identifiers, schema validation, controlled vocabularies, and links to model, data, and assay artifacts. A third phase could test cross-platform exchange, role-specific interfaces, append-only audit histories, and rules for inheritance, recalibration, retirement, or reissue after model and data updates. Operational progression should be conditional on demonstrated usability, semantic preservation, security, and scientific benefit. Research should also examine whether layered presentation can reduce cognitive burden without hiding unresolved evidence. No stage should be interpreted as regulatory endorsement, universal deployment readiness, or authorization to replace accountable human review.

Conclusion

The molecular prediction evidence passport is proposed as a versioned evidence object that keeps a molecular prediction attached to the conditions under which it was produced, interpreted, and used. Its original contribution is to connect prediction identity, confidence, calibration, applicability, provenance, model and data versioning, uncertainty sources, intended decision consequences, experimental feedback, interoperability, and audit history without collapsing them into a single performance or trust score. The construct addresses the accountability gap created when an output travels farther than its evidence and when benchmark success, explanation, plausibility, or internal approval is mistaken for experimental or therapeutic validation. Its

purpose is not to certify predictions or automate pharmaceutical decisions, but to make the relevant evidence state and its limitations inspectable at the moment of use and reconstructable afterward. Whether this architecture improves scientific decision-making remains an empirical question requiring schema development, prospective evaluation, human-factors testing, and cross-system interoperability studies. Until such evidence exists, the passport should be treated as a bounded prediction-governance proposal whose value is conditional on data suitability, semantic clarity, expert oversight, and disciplined respect for the boundary between computational output and experimental knowledge.

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. 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
3. Bender A, Cortes-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
4. Bender A, Cortes-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
5. 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
6. 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
7. 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
8. 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
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. Walters WP, Murcko MA. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol. 2020;38(2):143-5. doi:10.1038/s41587-020-0418-2
11. 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
12. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y
13. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368. doi:10.1136/bmj.l6927
14. Wise J, Grebe de Barron A, Splendiani A, Balali-Mood B, Vasant D, Little E, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today. 2019;24(4):933-8. doi:10.1016/j.drudis.2019.01.008
15. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, et al. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol. 2019;37(4):358-67. doi:10.1038/s41587-019-0080-8
16. 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
17. 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
18. Soleimany AP, Amini A, Goldman S, Rus D, Bhatia SN, Coley CW. Evidential deep learning for guided molecular property prediction and discovery. ACS Cent Sci. 2021;7(8):1356-67. doi:10.1021/acscentsci.1c00546
19. 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

20. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell. 2019;1(1):20-3. doi:10.1038/s42256-018-0004-1
21. Kompa B, Snoek J, Beam AL. Second opinion needed: Communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4:4. doi:10.1038/s41746-020-00367-3
22. 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
23. 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
24. Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, Norinder U, et al. Deep Docking: A deep learning platform for augmentation of structure based drug discovery. ACS Cent Sci. 2020;6(6):939-49. doi:10.1021/acscentsci.0c00229
25. 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
26. 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. doi:10.1016/j.cell.2020.01.021
27. Graff DE, Shakhnovich EI, Coley CW. Accelerating high-throughput virtual screening through molecular pool-based active learning. Chem Sci. 2021;12(22):7866-81. doi:10.1039/D0SC06805E
28. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020;18:1414-28. doi:10.1016/j.csbj.2020.05.017
29. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Felix E, et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1). doi:10.1093/nar/gky1075
30. Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, et al. Open Targets Platform: Supporting systematic drug-target identification and prioritisation. Nucleic Acids Res. 2021;49(D1). doi:10.1093/nar/gkaa1027
31. 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). doi:10.1093/bib/bbac404