Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2025, 14(3):184-194

https://doi.org/10.51847/wZ1d8zMJUu



Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Chemical Identity Before Molecular Prediction: A Pharmacoinformatics Framework for Resolving Stereochemistry, Salt Forms, Tautomerism, and Duplicate Records in Pharmaceutical Knowledge Bases

Marta Nowak[1]*, Hassan Al-Rashid[2], Aisha Bello[3]

[1]Department of Pharmaceutical Chemistry, Warsaw University of Technology, Warsaw, Poland.
[2]Department of Pharmaceutical Sciences, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.
[3]Department of Pharmaceutical Sciences, University of Lagos, Lagos, Nigeria.

***Email:** marta.nowak@pw.edu.pl*

ABSTRACT

Pharmaceutical knowledge bases frequently join structural records, substance descriptions and bioactivity observations through molecular identifiers whose normalization rules are rarely identical. A canonical structure can therefore provide reliable computational matching while obscuring distinctions relevant to stereochemistry, tautomerism, salt composition and the material actually tested. This conceptual and methodological analysis examines published molecular-curation procedures, chemical databases and predictive-data studies, and reconceptualizes duplicate detection as an endpoint-conditioned relation between provenance-bearing records. It distinguishes graph representation, assigned chemical state, material composition, observed assay entity and prediction unit, and proposes auditable rules for preserving these distinctions during integration. Competing explanations include assay heterogeneity, identifier errors and data-partition leakage. The resulting decision specification is offered for prospective expert adjudication and controlled modeling comparisons; its accuracy and utility have not been empirically validated. Chemical identity must be established at the level required by the intended pharmaceutical inference before molecular records are merged or prediction labels pooled.

Key words: *Chemical identity, Pharmacoinformatics, Stereochemistry, Tautomerism, Pharmaceutical salts, Provenance-aware deduplication*

Received: 18 July 2025; **Revised:** 21 August 2025; **Accepted:** 28 August 2025

INTRODUCTION

Chemical registration supplies the apparently simple premise that observations referring to the same molecule may be joined. Yet molecular data processing performs different operations under the shared label of standardization: parsing an input structure, correcting an invalid valence, selecting a canonical tautomer, extracting a neutral parent and associating a tested salt with an active ingredient. Their resulting identities are policy-dependent [1]. The pharmaceutical question is therefore which entity is being declared identical, for which inference, under which retained conditions.

The issue becomes consequential in predictive modeling. A training row inherits both a molecular encoding and an experimental outcome; alteration of either changes the scientific object the model is asked to learn. Comparative work on curation and confidence levels demonstrates sensitivity of compound predictions to data preparation [2]. Other causes include source quality, label choice and validation partitions.

© 2025 The International Journal of Pharmaceutical Research and Allied Sciences (IJPRAS). **Open Access** - This article is under the **CC BY NC SA** license (https://creativecommons.org/licenses/by-nc-sa/4.0).

Cross-source integration adds a second ambiguity. Data consistency checks in molecular-property prediction identify problems that remain even after compounds have been mapped into a common representational space [3]. An identical canonical string cannot certify that two reported activity values share endpoint definitions, experimental protocols or units. Different source strings can still support a task-specific relationship. A principled decision therefore needs more than a universal molecular-key equality test.

Form selection in drug development makes the boundary tangible: an active moiety may occur in distinct salts and solid forms whose selection requires material and developability judgments [4]. This article develops a nonempirical construct reconceptualization: record duplication is treated as a typed, evidence-qualified assertion about the unit of a particular downstream decision. Published methods motivate the distinctions; the proposed decision semantics remain testable propositions, not results from an executed validation study.

Chemical identity as a foundation for pharmaceutical data interpretation

A molecule may be identified at several resolutions. A connectivity description expresses atom-bond relationships; a stereochemical assignment adds configuration at specified centers; a tautomer convention chooses a depiction within an interconverting family; and a substance description may specify counterions, solvates or relative component amounts. None of these fields independently identifies the sample introduced into an instrument. Work predicting natural-product stereochemistry illustrates the usefulness of computational inference where assignment is incomplete [5]. A predicted assignment is not experimental proof of identity.

Stereochemistry supplies a practical test of representation granularity. Stereochemically aware bioactivity descriptors were developed because configurations that share connectivity may encode different target-interaction information [6]. For stereoselective endpoints, collapsing isomer records can hide the modeled distinction. For an achiral bulk property under specified circumstances, a coarser representation may be defensible. The correct unit cannot be determined without reference to the requested endpoint and evidence about the tested substance.

Chirality descriptors also depend on the encoding and learning system through which configuration reaches a model [7]. A model's ability to separate isomeric strings is distinct from evidence about enantiomeric purity, absolute configuration or a changing stereoisomeric mixture. Unspecified stereocenters, reported racemates and single stereoisomers are not interchangeable observations.

Layered International Chemical Identifier (InChI) representations enable systematic comparison of specified molecular properties, with important improvements to implementation reliability [8]. But an identifier resolves the conventions within its own layers, rather than experimentally establishing the identity of an administered or tested material. Equality warrants comparison only at declared layers and conventions.

PubChem's published standardization account shows that normalization rules govern charges, stereochemistry and molecular presentation [9]. This supports distinguishing three objects at the start of a pharmaceutical record: the source-submitted description, the rule-generated structural representation, and the material or measurement assertion supplied by the data source. A prediction entity is then an explicitly chosen projection from these objects. Computational convenience cannot certify dose or clinical equivalence.

Existing evidence on molecular representation and chemical record standardization

Automated quantitative structure–activity relationship workflows seek reproducible inputs through rule-based standardization. A recent open-source QSAR-ready procedure exemplifies operations that regularize molecular structures for modeling [10]. Consistent inputs can discard intact composition or ionization information.

The RDKit-based ChEMBL pipeline formalizes checking, standardization and extraction of a parent structure as separable processing operations [11]. This staged design is important because the route from a submitted source structure to a preferred parent is itself evidence about what has changed. Retaining both views enables subsequent review and task-specific use.

The PubChem resource distinguishes source-associated substance information from integrated compound-level representations and identifiers [12]. Linkage does not identify whether a record describes salt, neutral graph or tested batch. Provenance and material identity remain separate questions even when the structure representation is valid.

DrugBank integrates drug, molecular and target-related knowledge for diverse pharmacological uses [13]. Its breadth is valuable for joining knowledge about active substances and therapeutic contexts. A drug entry and a cheminformatics parent may nevertheless refer to different levels of pharmaceutical description. A cross-reference cannot certify equivalent formulations.

ChEMBL associates chemical structures with curated assays and activity records [14]. The resulting linkage is indispensable for medicinal-chemistry analysis but introduces two independent decisions: which structures should count as the same prediction input, and which reported endpoints are comparable. Graph agreement cannot correct assay-design or unit differences.

A retrospective account of ChEMBL describes expansion of the resource and its role in cheminformatics and drug discovery [15]. A single compound view may combine publications, deposits and annotation versions. The appropriate response is a reversible correspondence among source records, not the presumption that every linked measurement is a replicate of one invariant chemical entity.

Recent work integrating artificial intelligence with manual bioassay annotation reinforces the distinct role of experimental metadata [16]. Chemical name resolution may be automated while mechanism, format or endpoint interpretation still needs specialized checking. Structure and assay mappings may warrant different confidence judgments.

ChEBI's continued reengineering illustrates the separate contribution of chemical ontology and maintained identifier relationships to interoperability [17]. An ontological relationship such as membership in a chemical class, or a synonym or cross-reference, should not be mistaken for the identity of two assayed preparations. Ontology linkage alone cannot establish assay interchangeability.

BindingDB documents protein–small-molecule binding observations within an increasingly FAIR-oriented knowledgebase [18]. Association of a ligand with a binding record is meaningful only alongside the measurement's target, assay format and reported endpoint. Standardized records still need separate measurement-comparability evidence. **Table 1** specifies where each representational operation can legitimately support a pharmaceutical inference, and where the inference must stop.

Table 1. Identity-bearing units retained and transformed during pharmaceutical structure curation

Unit available to the analyst	Typical representational operation	What the record can support	Conditional inference requiring further evidence	Proposed preservation rule
Submitted structure or substance record	Parse, validate, repair annotations	A source supplied a particular graph or composition statement	That the entered graph represents the tested vial	Preserve source string, source identifier and edit history
Standardized connectivity graph	Normalize aromaticity, charge and valence	Graph comparison under stated software rules	Identical stereochemical substance or assay exposure	Store canonicalizer and transformation version
Stereochemically specified structure	Retain or encode configuration	Configurational distinction in representation	Absolute assignment or sample stereopurity	Keep explicit unknown, racemate and verified states separate
Parent extracted from salt/solvate	Separate fragments or discard counterions	A parent-relationship view useful for some searches	Same material stability, dissolution or delivered dose	Link intact form and parent without deleting components
Tautomer-normalized structure	Apply a chosen tautomer rule set	Comparison within a declared representational convention	The same equilibrium distribution in an experiment	Retain enumerated alternatives and rule scope
Assay-linked compound record	Join structure to endpoint metadata	A reported measurement associated with a chemical description	Cross-study label exchangeability	Keep source, assay, conditions and units with each observation

Table note: Evidence basis: the representation and knowledgebase practices described in this section. The final-column requirements are original decision proposals and have not been validated as a universal curation standard. No row asserts sample-level or therapeutic equivalence.

Problems of chemical equivalence, ambiguity, and duplicate identification

Tautomeric conflicts expose how a seemingly straightforward duplicate designation depends on the equivalence rule used. A published comparison of forty databases identified pairs or sets of records related under specified tautomer transformations [19]. Such conflicts do not establish physical speciation or justify averaging bioactivity.

A broader computational treatment of tautomerism describes the rule-design difficulty, including the treatment of distinct tautomer classes in identifiers [20]. Equivalence can become broader as more transformations are admitted; thus it is essential to distinguish a canonical representative from the chemical ensemble it is intended to index. Unspecified rules make duplicate counts difficult to reproduce.

The curated Tautomer Database helps reveal the role of published observations and experimental conditions in tautomer interpretation [21]. A reported interconversion or structure assignment is embedded in analytical method, solvent and related chemical environment. Structural membership in a tautomer family therefore provides a candidate link. A family label does not specify its distribution under a new assay.

Mixtures illustrate a different limit. A machine-readable scheme for mixed-substance composition records components and their relationships, information that disappears when a single parent graph replaces the substance [22]. This loss may be unimportant for a connectivity search but decisive for potency denominators, component-specific exposure, formulation composition and claims about reproducible material. The compared unit must be declared.

Canonical tautomer selection can also reflect decision rules informed by quantum chemistry [23]. The ranking or chosen depiction answers a clearly bounded computational task; solvent, proton activity, substituent environment and observation time may influence an experimental ensemble differently. Indexing resolution cannot decide which tautomer binds a target.

The canSAR chemical registration pipeline shows the utility of hierarchical relationships among standardized compounds and related forms [24]. A3 extends that representational insight with an original distinction: *exact recorded duplicate* means redundant assertions of a specified record at a source-and-version level; *representation-equivalent* means agreement after explicit normalization; *chemically related* covers justified parent, stereo or tautomer relations; and *measurement-comparable* demands independently documented assay and material conditions. These relations are noninterchangeable. **Figure 1** shows the resulting nested identity graph and its scientific limits.

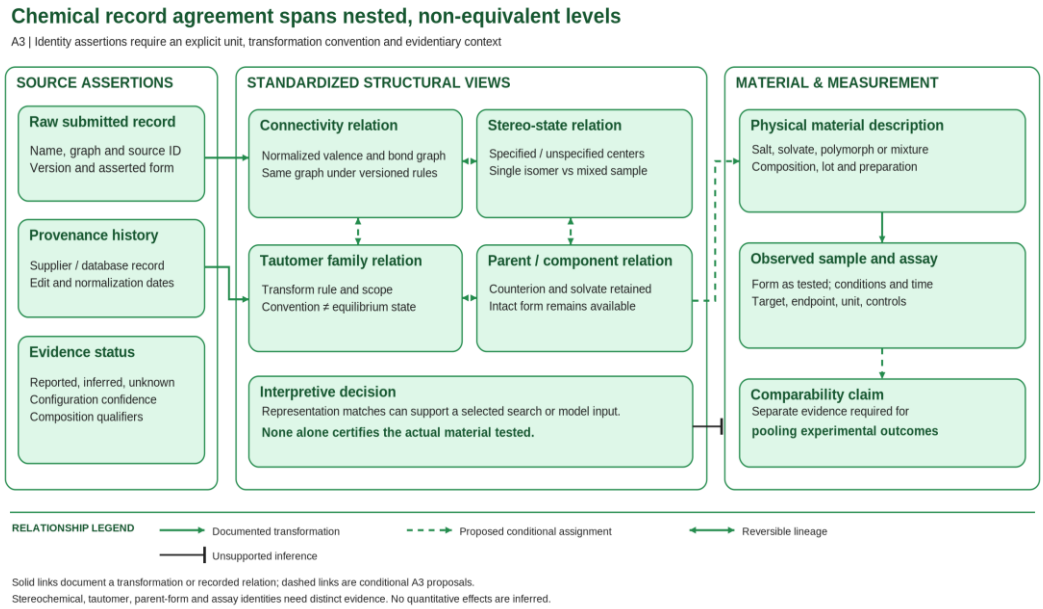


Figure 1. Chemical record agreement occupies nested levels that do not certify material or assay equivalence

Figure 1 caption/legend: The nested graph is read from source-record evidence toward progressively selected structural relations; cross-cutting material-state and assay-observation regions remain separate from pure structure normalization. A solid arrow denotes a documented representational transformation or published relation; a dashed arrow denotes a proposed conditional assignment; double-headed links denote reversible record navigation. A blunt-ended link marks a prohibited inference without independent material or measurement evidence. Distinct enclosure contours separate source record, connectivity, stereochemical assignment, tautomer family, intact salt/mixture and measured sample. The diagram encodes conceptual dependencies, not reaction rates or measured equivalence. “Canonical” denotes a convention-dependent normalized representation; a “parent” is a selected substructure view and need not represent the original material.

Figure 1 alt text: Wide nested chemical-identity graph. An outer source-record enclosure contains a standardized connectivity region, branching to stereo-specified and tautomer-related structural views; the intact salt or mixture form remains a sibling rather than a subordinate identity. A separate assay/sample zone crosses these branches. Solid transformation arrows lead between retained representations, dashed conditional lines connect them to possible measurement comparability, and blunt bars stop unsupported claims that shared canonical structure proves an identical tested preparation. A compact legend explains status-coded line types and hierarchy.

Unresolved challenges for molecular prediction and database integration

Even a carefully stratified identity relation leaves a measurement problem. Matched-pair analysis of inter-assay variability demonstrates why predictions and comparisons must be interpreted with respect to the experimental observation process [25]. A pair of molecules can have reproducible relative behavior under one collection of assays while absolute activity values from heterogeneous methods remain difficult to pool. Assay misclassification can mimic an identity error.

Comparative benchmarking of physicochemical and toxicokinetic predictors underscores dependence on endpoint definitions and model applicability [26]. A structure may be usable for one prediction target while insufficient for another requiring ionization, stereochemical purity or the intact administered form. A3 therefore treats the model input unit as a stated analytical decision, not an invariant consequence of a name resolver. Its predictive consequences remain untested.

The contemporary literature on data-driven new approach methodologies emphasizes applicability-domain reasoning and endpoint-sensitive interpretation [27]. Extrapolation beyond the training conditions may fail despite exact structure matching. The relationship between an in-silico descriptor and a pharmaceutical decision consequently needs an explicit chain of inference linking the descriptor, assay endpoint and material conditions relevant to deployment.

Automated, interpretable chemical grouping shows that clustering and descriptor-defined neighborhoods can serve useful comparisons [28]. Similarity groups are still distinct from duplicate groups: two dissimilar-looking inputs may be equivalent under one normalization, while nearby descriptor vectors can represent distinct stereoisomers. Conflating a grouping for model explanation with a claim of substance identity would corrupt error analysis and obscure applicability limits.

The Chemical Checker integration protocol demonstrates the extensive processing required to integrate heterogeneous bioactivity data into a compound-centered computational universe [29]. Such integration supports broad comparative signatures. It also poses a question upstream of modeling: did the signatures originate from observations that share an experimentally meaningful unit? Keep each signature's assay and preprocessing provenance.

Chemical biology resources such as the European Chemical Biology Database broaden available assay evidence, including screening contexts [30]. Divergent labels may reflect negative observations, assay interference, context-specific mechanisms or recording mistakes, none automatically corrected by a shared compound key. **Table 2** distinguishes rival causes of apparent discordance.

Table 2. Rival causes of apparent chemical-record discordance during predictive-data integration

Observed collision or disagreement	Candidate source of discordance	Evidence needed for adjudication	Risk if records are pooled	Status of A3 resolution action
Equal canonical connectivity, different activity	Assay format or biological target context	Assay identifiers, target, units, controls and conditions	Mixed-endpoint label	Proposed: assess measurement comparability independently
Different tautomer strings, linked transformations	Distinct canonicalization or tautomer conventions	Rule set, software version and experimental medium	False duplicate or false distinction	Proposed: preserve both strings and qualified family link
Shared parent, different counterions	Salt-form or composition variation	Intact sample form, constituent information, preparation	Loss of form-sensitive covariates	Proposed: use parent key only for justified tasks
Same graph, unspecified versus	Missing configuration or	Assigned centers and stereochemical analytical evidence	False stereochemical sameness	Proposed: defer exact-state merge pending evidence

explicit stereochemistry	mixture composition			
Similar chemistry, discordant model errors	Domain shift, split leakage or noisy labels	Source-held-out partitions, confidence tiers, endpoint consistency	Invalid generalization estimate	Proposed: adjudicate with paired sensitivity tests

Table note: Rows distinguish competing interpretations, not frequencies or established causation. Proposed actions specify what additional evidence must be retained before any endpoint-specific merge; the published studies supply the rationale for investigating these mechanisms.

Development of the proposed chemical-identity resolution framework

Identifier reconciliation begins with competing evidence, not with forced consensus. MoleculeResolver demonstrates the practicality of crosschecking structure resolution across services when names and identifiers are ambiguous [31]. Provider agreement can reflect shared upstream mistakes. Under the present specification, each resolution candidate retains the queried name, provider response, software timestamp or version where available, and a confidence statement whose scope is limited to that mapping task.

Integration in canSAR shows why a structure may occupy several legitimate positions within drug-discovery knowledge resources [32]. The proposed decision distinguishes the *record object* (what a named provider asserted), the *representation object* (a normalized encoding under named rules), the *chemical-state object* (configured graph, tautomer relation, ionic or material composition) and the *measurement object* (source-linked observation with experimental context). Prediction entities are selected for specific endpoints.

Cheminformatics microservices support reproducible manipulation and exchange of molecular descriptions [33]. Versioning need not imply universal equivalence. A proposed resolution decision should retain five fields: relation type, evidence basis, transformation provenance, unresolved state and permitted downstream uses. This is a conceptual data contract, not an assertion that a new software system has been built or evaluated.

Resolution should be reversible at the evidence level. If two salt records are joined through a shared parent, both original formulae and component descriptions remain accessible; if stereochemistry is uncertain, the unknown status is represented explicitly. Hierarchical registration makes this kind of parent–child navigation technically intelligible [24]. Reversibility refers to evidence tracing, not physical chemical recovery.

A binary duplicate flag would suppress the circumstances in which rival views disagree. Intra-database tautomer conflicts show that the same records can be related differently under declared rules [19]. A3 accordingly assigns resolution outcomes as *exact-duplicate assertion*, *representation match*, *state- or parent-related*, *measurement-comparable* or *unresolved*, with each supported by separately stored provenance. Task-specific coarsening must retain a traceable alternative view. **Figure 2** maps this network of assertions and the controlled routes by which a prediction dataset inherits them.

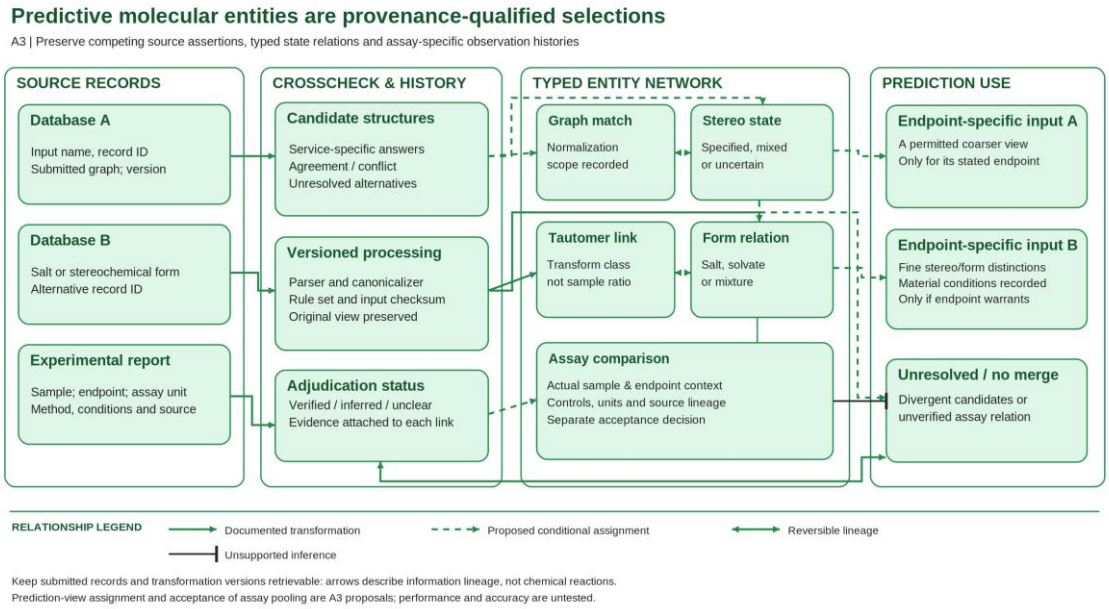


Figure 2. A prediction entity inherits typed chemical relations and source-level evidence without erasing discordant records

Figure 2 caption/legend: The provenance-based entity network separates source assertions, cross-service resolution candidates, versioned normalization operations, chemical-state relations, experimental observations and endpoint-specific model inputs. Solid arrows mark recorded source or transformation links; dashed arrows mark proposed conditional adjudications; two-headed links show reversible navigation; blunt-ended connectors forbid pooling when required material or assay metadata are absent. A fork denotes a genuinely unresolved alternative, not an empirically estimated probability. The network permits multiple prediction views of related records while preserving original identifiers, method versions and incompatible assay labels. It is a methodological proposal requiring expert evaluation and controlled testing.

Figure 2 alt text: Wide asymmetric network with several incoming database record nodes on the left feeding named cross-provider candidate and standardization zones. The center contains separate connectivity, stereo, tautomer and salt/mixture relation enclosures with archived original-record return links. Along the lower edge, assay observations enter a measurement-comparability lane. On the right two endpoint-specific predictive entities are derived using dashed, conditional assignments; ambiguous cases end in an unresolved branch. A legend distinguishes solid documented relations, dashed proposed decisions, bidirectional provenance return and blunt prohibition links.

Implications for pharmaceutical knowledge bases and predictive modelling

A molecular-learning benchmark usually presupposes that each input represents a stable prediction unit. MoleculeNet makes explicit the importance of endpoint definition, dataset construction and evaluation splits [34]. Chemical registration adds an upstream dependency: identical structures may index observations from distinct samples, while related but structurally nonidentical inputs may be duplicates under a deliberately coarser task definition. Each dataset view should disclose its identity relation. An exact record repeat differs from an independently replicated measurement even when both have the same recorded molecular structure.

Record integration could retain the intact source entity, a versioned normalization product, qualified chemical-state relationships and each measurement in a separately addressable form. Evidence of data-integration inconsistencies [3] supports examining these layers independently. As an original operational proposal, pooling would require both an explicitly chosen identity relation and justified comparability of endpoints. Unresolved stereochemistry or mixture composition would remain queryable attributes, including when a parent structure supplies a useful search index. This enables reproducible alternative dataset views.

For model evaluation, the distinction matters differently across tasks. Benchmarks of physicochemical and toxicokinetic prediction [26] show that prediction quality is endpoint-dependent. Connectivity may be insufficient when counterions, configuration or assay conditions alter observations. **Table 3** assigns proposed actions to the object that a modeling decision actually changes; these actions have not been shown to improve predictive performance.

Table 3. Proposed assignment of pharmaceutical record relations to prediction-dataset actions

Identity circumstance	Prediction unit at issue	Evidence to preserve	Proposed dataset action	Decision that remains unsupported
Repeated source assertion	Source record and version	Original identifier, deposition history	Mark provenance-level repeat; retain source trace	Independent replicated experiment
Canonically equal graphs	Structure encoding	Normalization rules and original strings	Share an index only at declared graph resolution	Same stereochemical sample
Configured versus unspecified stereo	Isomer or mixture	Assignment method and stereopurity	Keep distinct or unresolved states	Isomer-specific label interchangeability
Tautomer-family relation	Interconverting chemical descriptions	Tautomer rules and medium where known	Link family members; retain recorded forms	Identical assay speciation
Parent shared by different salts	Intact preparation and active moiety	Counterion, stoichiometry, form	Permit parent view; retain intact-form covariates	Same dissolution or exposure
Multiple bioactivity observations	Measured endpoint	Protocol, target, units, conditions	Assess label compatibility separately	Poolability from a shared compound key

Table note:* Every action in the fourth column is a proposed rule for prospective evaluation. A relation among representations, materials or assays is not inferred solely from a shared database identifier. No reported effect size or validation outcome is asserted.

Illustrative identity-resolution cases and proposed decision principles

Consider two records depicting opposite absolute configurations at the same stereogenic center. A shared connectivity key cannot justify pooling stereoselective labels. If one configuration was computationally assigned while the other was experimentally established, an identity decision must retain their different evidence statuses. Published work on machine-learning-assisted stereochemical assignment establishes a relevant inference task [5]; the case here is hypothetical, and does not demonstrate how a particular drug behaves.

Now consider records represented by different tautomer drawings, with a documented rule mapping both into one canonical family. The appropriate relation is family membership under that rule. Chemical equilibrium in a specified medium remains an independent question, as illustrated by curated tautomer evidence [21]. A predicted binding endpoint may need a state-sensitive representation, while a catalog search may use a family index. Neither establishes a measured tautomer ratio.

A third hypothetical pair describes the same neutral active moiety in distinct salt preparations. A parent relation aids searching but can conceal composition or concentration denominators. Structured representations of mixtures preserve component descriptions [22]. This case motivates keeping the tested material as a separate identity object until the assay report establishes which preparation was used.

Finally, two databases may provide conflicting tautomer assignments for one named entry. Quantum-chemistry-informed standardization rules can help choose a consistent depiction for registration [23]. They cannot determine whether both providers drew on the same original experiment or whether a deposited activity label was misassociated. Use versioned rules for representation and provenance for observations. Cases that lack either remain explicitly unresolved.

Discussion and future research priorities

A chemical-identity policy can improve apparent modeling results for reasons unrelated to chemistry. Removing repeated entries may prevent train–test leakage; curation may also exclude low-confidence labels or alter the evaluation population. Published investigations of curation effects [2] therefore motivate controlled comparisons, with identical algorithms, endpoint definitions and splits across competing identity policies. Observed gains require causal attribution.

A decisive prospective test would start with blinded expert classification of source-record pairs into exact repeat, representation match, qualified chemical relation, measurement-compatible observation or unresolved. Data-integration findings [3] motivate a second, independent annotation of assay compatibility. Reviewers should be blinded to policy-specific predictions. Disagreement should be reported by relation type, with an adjudication record, instead of hidden through a majority merge.

The nearest rival account is that many alleged chemical-identity conflicts are artifacts of canonicalization conventions, not differences in test materials. Documented tautomer conflicts [19] make this rival plausible. A discriminating study would separately vary the normalization rules, the decision to preserve chemical-state features and the assay-label harmonization policy. If label harmonization suffices, identity annotation may add little.

Testing should compare source-held-out, scaffold-held-out and temporally separated evaluations where feasible, keeping the split strategy constant within each contrast. The matched-pair literature illustrates why assay variability may persist despite careful structural comparisons [25]. Record-level adjudication accuracy, unresolved-case frequency, endpoint-specific prediction error and calibration would be prespecified outcomes; no direction or magnitude is predicted here. An acceptable result could show that richer identity annotation changes auditability without improving prediction.

Limitations and framework boundaries

The account concerns the semantics of recorded chemical assertions and their use in pharmaceutical prediction. Solid-state form selection, preparation history and dosing information cannot be reconstructed from a molecular graph. The drug-development literature on form selection [4] establishes the importance of material context, but A3 supplies no experiments demonstrating when a particular counterion or polymorph changes a given endpoint. Preparation-specific evidence remains necessary.

Canonical encodings are constrained by supported chemical classes, parser settings and software versions. The reliability of an InChI implementation [8] provides a useful interoperability foundation without validating every source assignment or verifying material identity. Incorrect source structures can be faithfully canonicalized; unresolved mixtures and uncertain configurations can remain unresolved after all reasonable transformations.

The advocated separation of original record, normalized representation and predictive view is an unvalidated conceptual specification. Existing automated standardization workflows [10] offer practical comparators, yet A3 has not measured annotation cost, inter-annotator agreement, computational overhead, retrieval quality or model generalization. Richer annotation may reduce coverage.

Finally, a knowledge base can preserve the assay metadata available to it but cannot supply missing controls, exposure measurements or inaccessible provenance. Diverse screening resources [30] are useful comparison contexts, although unequal endpoint definitions and biological systems restrict transport across experiments. No claim of clinical interchangeability, pharmacological mechanism or patient outcome follows from the chemical-registration relations considered here.

CONCLUSION

For pharmaceutical prediction, chemical identity is a question about the entity to which a particular observation and decision refer. Structural normalization establishes reproducible representational relationships; it cannot by itself resolve stereochemical composition, tautomeric state in an assay, intact salt form or measurement comparability. The proposed response is to preserve source assertions, attach typed and versioned identity relations, and select prediction units with explicit endpoint qualifications. This preserves the ability to distinguish genuine repeats from related forms and conflicting observations. Its practical value remains to be tested against simpler curation policies under matched measurement and evaluation conditions.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. Esaki T, Ikeda K. Data curation in cheminformatics: Importance and implementation. J Cheminform. 2026;18(1):43. doi:10.1186/s13321-026-01174-w
2. Xerxa E, Vogt M, Bajorath J. Influence of data curation and confidence levels on compound predictions using machine learning models. J Chem Inf Model. 2024;64(24):9341-9. doi:10.1021/acs.jcim.4c01573
3. Parrondo-Pizarro R, Menestrina L, Garcia-Serna R, Fernández-Torras A, Mestres J. Enhancing molecular property prediction through data integration and consistency assessment. J Cheminform. 2025;17(1):163. doi:10.1186/s13321-025-01103-3
4. Reid DL, Faul MM, López-Mejías V, Agarwal P, Bergauer M, Blue LE, et al. Application of an interdisciplinary approach to form selection in drug development. Org Process Res Dev. 2025;29(2):237-54. doi:10.1021/acs.oprd.4c00320
5. Orsi M, Reymond JL. Assigning the stereochemistry of natural products by machine learning. J Cheminform. 2026;18(1):76. doi:10.1186/s13321-026-01205-6
6. Comajuncosa-Creus A, Lenes A, Sánchez-Palomino M, Dalton D, Aloy P. Stereochemically-aware bioactivity descriptors for uncharacterized chemical compounds. J Cheminform. 2024;16(1):70. doi:10.1186/s13321-024-00867-4
7. Baimacheva N, Gao X, Aires-de-Sousa J. Evaluation of chirality descriptors derived from SMILES heteroencoders. J Cheminform. 2025;17(1):137. doi:10.1186/s13321-025-01080-7
8. Goodman JM, Pletnev I, Thiessen P, Bolton E, Heller SR. InChI version 1.06: Now more than 99.99 % reliable. J Cheminform. 2021;13(1):40. doi:10.1186/s13321-021-00517-z
9. Hähnke VD, Kim S, Bolton EE. PubChem chemical structure standardization. J Cheminform. 2018;10(1):36. doi:10.1186/s13321-018-0293-8
10. Mansouri K, Moreira-Filho JT, Lowe CN, Charest N, Martin T, Tkachenko V, et al. Free and open-source QSAR-ready workflow for automated standardization of chemical structures in support of QSAR modeling. J Cheminform. 2024;16(1):19. doi:10.1186/s13321-024-00814-3

11. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, et al. An open source chemical structure curation pipeline using RDKit. J Cheminform. 2020;12(1):51. doi:10.1186/s13321-020-00456-1
12. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2025 update. Nucleic Acids Res. 2025;53(D1):D1516-25. doi:10.1093/nar/gkae1059
13. Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, et al. DrugBank 6.0: The DrugBank knowledgebase for 2024. Nucleic Acids Res. 2024;52(D1):D1265-75. doi:10.1093/nar/gkad976
14. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1):D930-40. doi:10.1093/nar/gky1075
15. Zdrazil B. Fifteen years of ChEMBL and its role in cheminformatics and drug discovery. J Cheminform. 2025;17(1):32. doi:10.1186/s13321-025-00963-z
16. Smit I, Adasme MF, Manners E, Corbett S, Bosc N, Do HMA, et al. Integrating artificial intelligence and manual curation to enhance bioassay annotations in ChEMBL. J Cheminform. 2026;18(1):24. doi:10.1186/s13321-026-01165-x
17. Malik A, Arsalan M, Moreno C, Mosquera J, Félix E, Kizilören T, et al. ChEBI: Re-engineered for a sustainable future. Nucleic Acids Res. 2026;54(D1):D1768-78. doi:10.1093/nar/gkaf1271
18. Liu T, Hwang L, Burley SK, Nitsche CI, Southan C, Walters WP, et al. BindingDB in 2024: A FAIR knowledgebase of protein-small molecule binding data. Nucleic Acids Res. 2025;53(D1):D1633-44. doi:10.1093/nar/gkae1075
19. Dhaked DK, Nicklaus MC. Tautomeric conflicts in forty small-molecule databases. J Chem Inf Model. 2024;64(19):7409-21. doi:10.1021/acs.jcim.4c00700
20. Dhaked DK, Ihlenfeldt WD, Patel H, Delannee V, Nicklaus MC. Toward a comprehensive treatment of tautomerism in chemoinformatics including in InChI V2. J Chem Inf Model. 2020;60(3):1253-75. doi:10.1021/acs.jcim.9b01080
21. Dhaked DK, Guasch L, Nicklaus MC. Tautomer database: A comprehensive resource for tautomerism analyses. J Chem Inf Model. 2020;60(3):1090-100. doi:10.1021/acs.jcim.9b01156
22. Clark AM, McEwen LR, Gedeck P, Bunin BA. Capturing mixture composition: An open machine-readable format for representing mixed substances. J Cheminform. 2019;11(1):33. doi:10.1186/s13321-019-0357-4
23. Baker CM, Kidley NJ, Papachristos K, Hotson M, Carson R, Gravestock D, et al. Tautomer standardization in chemical databases: Deriving business rules from quantum chemistry. J Chem Inf Model. 2020;60(8):3781-91. doi:10.1021/acs.jcim.0c00232
24. Dolciami D, Villasclaras-Fernandez E, Kannas C, Meniconi M, Al-Lazikani B, Antolin AA. canSAR chemistry registration and standardization pipeline. J Cheminform. 2022;14(1):28. doi:10.1186/s13321-022-00606-7
25. Nelen J, Pérez-Sánchez H, De Winter H, Van Rompaey D. Matched pairs demonstrate robustness against inter-assay variability. J Cheminform. 2025;17(1):8. doi:10.1186/s13321-025-00956-y
26. Gadaleta D, Serrano-Candelas E, Ortega-Vallbona R, Colombo E, Garcia de Lomana M, Biava G, et al. Comprehensive benchmarking of computational tools for predicting toxicokinetic and physicochemical properties of chemicals. J Cheminform. 2024;16(1):145. doi:10.1186/s13321-024-00931-z
27. Piir G, Sild S, Tcheremenskaia O, Di Consiglio E, Henriksson J, Gajewicz-Skretna A, et al. Perspective on applicability of data-driven machine learning computational new approach methodologies for hazard identification in chemicals risk assessment. J Cheminform. 2026;18(1):57. doi:10.1186/s13321-026-01184-8
28. Moreira-Filho JT, Ranganath D, Conway M, Schmitt C, Kleinstreuer N, Mansouri K. Democratizing cheminformatics: Interpretable chemical grouping using an automated KNIME workflow. J Cheminform. 2024;16(1):101. doi:10.1186/s13321-024-00894-1
29. Comajuncosa-Creus A, Bertoni M, Duran-Frigola M, Fernández-Torras A, Guitart-Pla O, Kurzawa N, et al. Integration of diverse bioactivity data into the Chemical Checker compound universe. Nat Protoc. 2025;20(11):3270-94. doi:10.1038/s41596-025-01167-3
30. Škuta C, Müller T, Voršilák M, Popr M, Epp T, Skopelitou KE, et al. ECBD: European chemical biology database. Nucleic Acids Res. 2025;53(D1):D1383-92. doi:10.1093/nar/gkae904
31. Müller S. How to crack a SMILES: Automatic crosschecked chemical structure resolution across multiple services using MoleculeResolver. J Cheminform. 2025;17(1):117. doi:10.1186/s13321-025-01064-7
32. Gingrich PW, Chitsazi R, Biswas A, Jiang C, Zhao L, Tym JE, et al. canSAR 2024—an update to the public drug discovery knowledgebase. Nucleic Acids Res. 2025;53(D1):D1287-94. doi:10.1093/nar/gkae1050

33. Rajan K, Chandrasekhar V, Sharma N, Kanakam SRS, Baensch F, Steinbeck C. Cheminformatics microservice V3: A web portal for chemical structure manipulation and analysis. J Cheminform. 2025;17(1):142. doi:10.1186/s13321-025-01094-1
34. 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