



Why Stereochemistry Must Be Explicitly Represented When ML Predicts Binding, Bioactivity, and Selectivity of Complex Natural Products

Hans Müller^{1*}, Anna Schmidt¹, Thomas Weber²

¹Department of Computational Stereochemistry and Molecular Recognition, Faculty of Pharmaceutical Sciences, University of Bonn, Bonn, Germany.

²Department of Machine Learning for Chiral Bioactivity, Faculty of Chemistry and Pharmacy, LMU Munich, Munich, Germany.

ABSTRACT

Machine-learning models for molecular prediction can use only chemical distinctions that survive data curation and molecular representation. For complex natural products, this constraint is especially important because stereochemical configuration may contribute to molecular identity, biological recognition, selectivity, and measured activity, yet it can be omitted, flattened, inconsistently resolved, or conflated with conformational information during data preparation. This conceptual article examines stereochemistry as an information-preservation problem rather than as a generic argument that three-dimensional representations are always superior. It integrates evidence from natural-product chemistry, stereosensitive molecular learning, three-dimensional fingerprints, geometry-aware graph models, conformer analysis, and cheminformatics pipelines. The central proposed contribution is a representation principle: when the target endpoint can plausibly depend on configuration, a model should preserve stereochemical identity explicitly, distinguish known from unknown or mixed stereochemical states, and be evaluated against stereo-blinded alternatives before claims of stereochemical sensitivity are accepted. Stereo-preserving two-dimensional graphs, geometry-aware representations, and conformer-aware approaches are therefore treated as complementary regimes rather than a hierarchy with one universally preferred method. The argument is bounded by task dependence, conformer uncertainty, assay noise, incomplete stereochemical annotation, and the possibility that some endpoints are effectively stereo-insensitive. Explicit representation is thus proposed as a condition for valid stereo-sensitive inference, not as evidence that stereochemistry will improve every benchmark or that three-dimensional modeling is always necessary. The framework requires endpoint-specific and prospective validation before deployment claims are justified.

Key Words: Stereochemistry, Natural products, Molecular machine learning, Chirality, Molecular representation, Bioactivity prediction

eIJPPR 2025; 15(1):86-95

HOW TO CITE THIS ARTICLE: Müller H, Schmidt A, Weber T. Why Stereochemistry Must Be Explicitly Represented When Machine Learning Predicts the Binding, Bioactivity, and Selectivity of Complex Natural Products. Int J Pharm Phytopharmacol Res. 2025;15(1):86-95. <https://doi.org/10.51847/NbwU239MLj>

INTRODUCTION

A molecular representation is not a neutral transcription of a molecule. It is an information-selection step that determines which chemical distinctions are available to a learning algorithm and which have already disappeared

before optimization begins. Contemporary molecular machine learning spans strings, fingerprints, graphs, learned embeddings, and three-dimensional encodings, each retaining a different subset of molecular information [1]. For stereochemistry, the consequence is unusually

Corresponding author: Hans Müller

Address: Department of Computational Stereochemistry and Molecular Recognition, Faculty of Pharmaceutical Sciences, University of Bonn, Bonn, Germany.

E-mail: ✉ hans.mueller@uni-bonn.de

Received: 30 October 2024; **Revised:** 20 January 2025; **Accepted:** 03 February 2025

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.



sharp: two structures can share atomic composition and connectivity while differing in spatial configuration, yet an encoding that collapses those configurations asks the model to treat chemically distinct inputs as computationally identical. More model capacity cannot repair a distinction that was removed upstream; it can only exploit other correlated features present in the training data.

Recent representation research further shows that molecular graphs need not be restricted to bare connectivity. Graph encodings can incorporate chemically structured information, while chirality-aware graph learning demonstrates that explicit chiral knowledge can be embedded for stereosensitive prediction tasks [2, 3]. These studies do not establish that every biological endpoint is stereosensitive, nor that explicit chirality will necessarily improve every benchmark. They establish the more limited but essential computational premise that stereochemical distinctions can be represented and learned when they are supplied to the model rather than erased beforehand.

That premise is especially pertinent to natural products. Natural-product drug discovery draws on chemically diverse and structurally elaborate molecules whose biological value is accompanied by challenges in characterization, dereplication, synthesis, and data integration [4]. Their complexity also includes ring systems, conformational constraints, unusual substitution patterns, macrocycles, and other features that cannot be reduced to stereochemistry alone. The argument developed here is therefore narrower: stereochemistry is one potentially pharmacologically consequential component of molecular identity that cannot be evaluated scientifically if preprocessing renders it invisible.

This article consequently treats stereo-sensitive prediction first as an information-integrity problem. Before asking whether a machine-learning system predicts binding, bioactivity, or selectivity accurately, one should ask whether stereoisomers remain distinguishable in the molecular records presented to that system. Explicit representation does not guarantee improved prediction. It establishes the minimum representational capacity required to discover, reject, or quantify stereochemical dependence rather than assuming such dependence away.

Why natural product stereochemistry is pharmacologically consequential

Natural products are sometimes discussed as though biosynthesis guarantees a single, fixed stereochemical state. That assumption is too simple. Racemic and scalemic natural products are documented, with chemical, biosynthetic, and post-biosynthetic processes capable of producing non-enantiopure material [5]. An analytical

study of cannabichromene in *Cannabis sativa* further showed that natural samples can differ in scalemicity and enantiomeric dominance [6]. For computational datasets, these observations mean that a record denoting “the molecule” may sometimes correspond to a specific stereoisomer, a nonracemic mixture, a racemate, or material whose enantiomeric composition depends on biological source or sample history.

The pharmacological relevance of that distinction must nevertheless be demonstrated endpoint by endpoint. A stereochemical label is not important merely because a stereogenic element exists. It becomes relevant when configuration changes the molecular state presented to a biological or analytical system. In one natural-product example, a chiral ^1H NMR method discriminated (R)- and (S)-cannabichromene in complex mixtures and paired stereochemical resolution with assessment of TRPA1 activity [7]. This does not establish a universal magnitude of enantiomeric divergence, but it demonstrates why unresolved stereochemical identity can complicate interpretation of functional labels.

The conceptual consequence is that stereochemistry belongs upstream of prediction. If assay values from distinct stereochemical states are merged under a representation that cannot distinguish them, the resulting inconsistency is chemically structured rather than ordinary random noise. Conversely, if stereoisomers behave similarly for a given endpoint, explicit representation still allows that similarity to be learned rather than imposed. The operative question is therefore not whether every natural-product stereoisomer differs pharmacologically, but whether the modeling workflow preserves the possibility of detecting when it does.

Where stereochemical information is lost in molecular data

Stereochemical information can disappear well before model fitting. Three-dimensional fingerprints illustrate the underlying issue: E3FP encodes spatial molecular neighborhoods and produces similarity relationships distinct from those obtained with conventional two-dimensional fingerprints, demonstrating that spatial organization constitutes an independent representational signal [8]. A two-dimensional representation is not intrinsically inadequate, but if its encoding omits stereochemical descriptors, configurationally distinct molecules may be mapped onto the same computational identity.

The same problem can arise inside otherwise sophisticated graph-learning workflows. Chemprop formalizes molecular featurization as a configurable stage preceding message passing, underscoring that the information available to a graph neural network depends on how atoms

and bonds are encoded [9]. Stereo retention is therefore a pipeline decision rather than an automatic consequence of using a graph architecture. Structure resolution, canonicalization, isomeric versus non-isomeric string export, fingerprint generation, and deduplication can each preserve or erase configuration depending on implementation. A model may then appear accurate because correlated scaffold or assay information predicts the label, while remaining fundamentally unable to distinguish stereoisomeric counterfactuals.

Cheminformatics experience outside property prediction makes the broader point explicit. RDChiral was developed because automated reaction-template extraction and application required dedicated logic for preserving, inverting, creating, or removing stereochemical assignments correctly [10]. Reaction informatics is not equivalent to bioactivity prediction, so this evidence should not be interpreted as a measured stereochemical error rate for natural-product datasets. It nevertheless demonstrates that generic molecular graph processing does not guarantee correct stereochemical semantics.

A stereo-sensitive workflow should therefore audit representational integrity across the complete path from source structure to model input. The critical failure is not simply “using 2D.” It is any transformation that converts distinct stereochemical states into the same model-visible object without making that invariance explicit, chemically justified, and appropriate to the intended endpoint.

Two-dimensional, three-dimensional, and conformer-aware representations

The appropriate response to stereochemical information loss is not to mandate three-dimensional modeling in every case. Natural-product fingerprint benchmarking shows that representation choice materially changes how natural-product chemical space is organized and can alter downstream bioactivity-model behavior [11]. This establishes representation sensitivity in natural-product space, but it does not identify chirality as the sole cause of those differences or establish a single representation as uniformly optimal.

Molecular representations that retain additional structural information can materially change downstream prediction behavior [11, 12]. Geometry-enhanced learning demonstrates that bond-angle and other geometric information can add predictive signal beyond simpler graph descriptions [12]. Yet geometry and stereochemistry are not synonyms. A stereo-preserving two-dimensional graph can distinguish configurations without requiring a generated conformer, while a three-dimensional model additionally becomes dependent on coordinate construction, conformational sampling, and—in structure-based applications—possible pose assumptions. Explicit stereochemistry therefore concerns preserving configurational identity; three-dimensional modeling concerns representing geometry, which is a related but separable modeling decision.

Conformer-aware modeling adds both information and uncertainty. Controlled analysis of conformation importance shows that greater three-dimensional detail is not automatically advantageous: usefulness varies with endpoint, representation, and conformer choice, and inappropriate conformations can fail to improve prediction [13]. This distinction is particularly important for flexible, macrocyclic, or conformationally rich natural products, for which a single generated geometry may poorly represent the relevant accessible ensemble. Treating one arbitrary conformer as molecular truth can therefore replace one form of information loss with another form of model misspecification.

The proposed distinction is functional rather than dimensional. Stereo-preserving two-dimensional encodings are defensible when configuration must remain distinguishable but conformational detail is not demonstrably necessary. Geometry-aware models become attractive when spatial relations are plausibly linked to the endpoint. Conformer ensembles become justified when multiple accessible geometries themselves form part of the prediction problem. **Table 1** aligns these representation mechanisms with the information each preserves, the evidence supporting its role, and the principal error introduced when that role is overextended.

Table 1. Mechanism–evidence alignment for stereo-preserving molecular representations in complex natural-product prediction

Representation mechanism	Information preserved	Evidence-aligned role	Principal failure if used uncritically
Stereo-aware 2D graph	Connectivity plus explicit configuration	Distinguishes stereoisomers without imposing one three-dimensional geometry	May omit conformational or interaction geometry required by the endpoint
Three-dimensional fingerprint	Local spatial arrangement	Adds spatial similarity information beyond conventional two-dimensional fingerprints	Depends on selected conformation; spatial similarity is not equivalent to biological equivalence
Geometry-enhanced graph	Topology plus geometric relations	Supplies spatial information while retaining graph-based learning	Added geometry may be irrelevant, approximate, or noisy for some endpoints

Conformer-aware representation	Multiple plausible molecular geometries	Represents conformational variability when that variability is endpoint-relevant	Sampling quality and conformer selection can dominate apparent benefit
--------------------------------	---	--	--

Note: This matrix is a manuscript-derived synthesis aligning established representation capabilities with a proposed endpoint-dependent use rule. It does not rank the four mechanisms universally or imply that increasing geometric detail necessarily improves stereo-sensitive prediction.

Stereochemistry in binding and bioactivity prediction

A model intended to make stereo-sensitive binding predictions must first possess a representation capable of distinguishing configurations. Geometry-complete perceptron networks illustrate this computational requirement because the same architecture can recognize molecular chirality and can also operate on protein–ligand binding-affinity tasks [14]. Those results demonstrate representational capability, not that chirality itself caused improved affinity prediction. A credible stereo-sensitive claim therefore requires more than competitive affinity metrics: the representation must distinguish the relevant stereochemical states, and evaluation must show that this distinction contributes information beyond scaffold, target, or dataset correlations.

Binding also depends on more than isolated ligand geometry. DynamicBind treats complex formation as a ligand-conditioned three-dimensional problem in which protein conformation can adapt to the ligand [15]. This reinforces a mechanistic reason stereochemistry may matter—configuration can constrain spatially feasible interactions—but predicted poses remain structural

hypotheses rather than measured affinities or functional responses. Protein flexibility, protonation, solvent effects, and alternative binding modes can all modify the apparent relationship between ligand configuration and the final assay label.

Structure-based affinity learning provides another representation regime. Pafnucy demonstrated that protein–ligand complexes can be encoded directly as three-dimensional grids for affinity prediction [16]. Such models can, in principle, preserve spatial distinctions lost in topology-only inputs, yet they introduce dependence on structure quality, poses, and database composition. Binding prediction should therefore be separated from functional bioactivity prediction: affinity concerns interaction strength under a defined model or assay, whereas functional bioactivity can include downstream signaling, permeability, metabolism, cytotoxicity, or other system-level processes.

Table 2 summarizes the manuscript-derived boundary conditions that determine when a stereo-sensitive binding claim is scientifically interpretable rather than merely representationally possible.

Table 2. Boundary conditions for interpreting stereo-sensitive binding and bioactivity predictions

Boundary condition	Why it matters	Required interpretation
Configuration is model-distinguishable	A stereo-blind representation cannot test configurational dependence	Stereo sensitivity requires explicit representational distinguishability
Protein geometry is uncertain or adaptable	Different ligand configurations may induce or prefer different complex geometries	Predicted configuration–binding relationships remain conditional on the structural model
Complex coordinates or poses are imperfect	Three-dimensional input can encode structural error as confidently as structural truth	Structure-based prediction is not equivalent to experimentally established binding mechanism
Functional activity exceeds binding alone	Cellular or organism-level responses add processes not represented by an affinity model	Binding, functional bioactivity, and downstream pharmacology require separate labels and validation

Note: The interpretive requirements are a proposed synthesis derived from the completed argument. The cited studies establish representational or structural capabilities; they do not validate a universal stereochemistry-to-binding pathway.

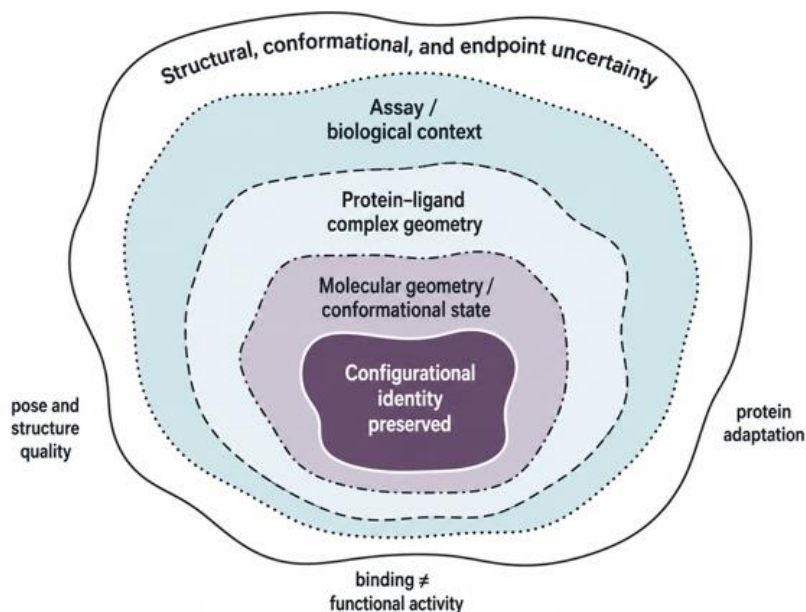


Figure 1. Model-visible stereochemistry across nested binding contexts

Figure 1 presents the resulting manuscript-derived view of stereochemical information as nested within progressively broader binding and bioactivity contexts rather than as an isolated molecular feature.

Selectivity, enantiomeric divergence, and off-target risk

Stereo-sensitive prediction becomes more demanding when the endpoint is selectivity rather than affinity to one target. Contemporary medicinal chemistry documents that stereoisomers can differ in target interaction, pharmacokinetics, efficacy, safety, or other development-relevant properties [17]. These possibilities do not mean that every stereocenter creates pharmacological divergence. They mean that collapsing configurations before modeling can prevent the model from determining whether such divergence exists.

The relevant stereochemical space also extends beyond classical tetrahedral centers. Configurationally stable atropisomers can behave as distinct molecular entities with different biological or developability properties [18]. Representation strategies should therefore declare what types of stereogenic information they preserve—tetrahedral, double-bond, axial, or other configurationally stable forms—and whether apparently distinct states interconvert on the timescale relevant to measurement.

A matched medicinal-chemistry example illustrates the selectivity problem more directly. Deliberate control of atropisomerism contributed to discovery of a selective and stable PI3K β inhibitor with differentiated properties within a related chemical series [19]. This is strong evidence that configuration can alter pharmacological behavior within a controlled scaffold context, but it does not provide a general effect size for natural products.

For machine learning, selectivity should consequently be treated as a structured prediction problem. A compound may be potent on one target yet insufficiently selective against homologs, may display functional selectivity downstream, or may generate exposure-dependent off-target effects. The proposed article logic therefore rejects a single undifferentiated “selective” label. Explicit stereochemical representation preserves the possibility that configuration contributes differently to each component while allowing the data to show when it does not.

Dataset annotation, enumeration, and label integrity

Stereo-sensitive modeling is only as reliable as the identity assigned to each training example. Natural-product resources necessarily integrate records from heterogeneous sources, and COCONUT 2.0 illustrates the substantial normalization, restructuring, and curation required to build an open natural-product collection [20]. Such curation is essential, but it cannot reconstruct experimentally unknown configuration. A chemically valid structure record can therefore still be insufficient for a stereo-sensitive prediction task if its absolute or relative configuration is unspecified.

Structure resolution is another vulnerability. MoleculeResolver demonstrates that mapping identifiers to chemical structures across public services can produce inconsistencies that justify explicit crosschecking [21]. In a stereo-sensitive workflow, resolving a name to a constitutionally correct but stereochemically incomplete structure is not a harmless formatting difference: it can merge distinct molecular states before modeling.

Natural-product provenance further complicates label interpretation. LOTUS explicitly links natural-product structures to organisms, occurrences, and source evidence

[22]. For stereochemistry, this suggests a necessary distinction among the structure asserted by a database, the material actually isolated, the method used to assign configuration, and the material used in the assay. These records need not always be identical in evidential strength. The proposed dataset principle is therefore to preserve label lineage rather than treating a canonical structure as self-sufficient. Known configuration, unknown stereochemistry, unresolved mixtures, scalemic material, enumerated hypothetical stereoisomers, and

experimentally isolated stereoisomers should remain distinguishable states. Enumeration can support controlled experiments or virtual counterfactuals, but an enumerated stereoisomer must never inherit an experimental bioactivity label merely because it shares connectivity with the measured material.

Table 3 converts these distinctions into a manuscript-derived validation-readiness matrix for stereo-sensitive natural-product datasets.

Table 3. Validation readiness of stereochemical records before natural-product machine learning

Dataset state	Minimum evidence needed	Permitted modeling use	Readiness consequence
Curated structure with assigned configuration	Traceable structure record and explicit stereo encoding	Stereo-sensitive training if the assay label maps to that state	Potentially ready after label-lineage checks
Structure resolved from names or identifiers	Crosschecked structural resolution	Use only after confirming stereochemical completeness	Conditional; resolution inconsistency can propagate upstream error
Natural-product occurrence record	Link among structure, biological source, and provenance	Supports provenance-aware dataset construction	Insufficient alone to establish experimentally measured configuration
Unknown, mixed, or enumerated stereo state	Explicit status rather than forced canonical assignment	Separate analysis, uncertainty handling, or proposed counterfactual evaluation	Not interchangeable with an experimentally defined stereoisomer

Note: “Readiness” denotes suitability for a stereo-sensitive modeling claim, not general database quality. The final row is a proposed data-governance distinction derived from the manuscript.

Evaluating stereo-sensitive models without shortcut learning

Explicit stereochemical features do not guarantee that a model uses them. Binding-affinity datasets can contain systematic biases that support apparently strong performance without transferable interaction learning; resolving identifiable bias can materially improve generalization [23]. A stereo-aware model should therefore be compared with an otherwise equivalent stereo-blinded control. If both perform similarly, the result may indicate that the endpoint is stereo-insensitive, that stereo labels are too sparse or noisy, or that the model is relying on correlated scaffold information.

Evaluation must also define what “generalization” means. Real-world molecular out-of-distribution analysis shows that different distribution shifts test different model weaknesses [24]. A model can consequently appear robust under one split and fragile under another; comparative OOD evaluation confirms that in-distribution performance does not guarantee robustness once chemical distribution shifts [24, 25]. For stereo-sensitive tasks, the article therefore proposes several non-equivalent tests: holding out stereoisomeric counterparts, holding out related scaffolds or series, separating source domains, and testing

external datasets where structure and assay provenance are independently curated.

Shortcut risk is particularly acute when train and test sets contain highly related analogues. Ligand-based benchmarks can reward memorization of dataset structure rather than transferable pharmacological relationships [26]. Stereoisomeric evaluation offers an unusually stringent counterfactual because connectivity can be held nearly constant while configuration changes. Yet such pairs are useful only when assay conditions and labels are comparable; otherwise the apparent stereo effect may be confounded by experimental context.

The decisive evaluation question is therefore not “does the stereo-aware model score higher?” but “does explicit stereochemical information improve a prediction that survives stereo-blinded ablation, matched-pair testing, leakage control, and relevant distribution shift?” Failure at any one stage does not prove stereochemistry is unimportant. It identifies where the evidence is insufficient to attribute predictive value specifically to configuration.

Table 4 summarizes the data-quality failures that can produce a misleading stereo-sensitive result and the consequence each creates for interpretation.

Table 4. Data-quality failure modes that can mimic stereo-sensitive model performance

Failure mode	Apparent result	Scientific consequence	Required control
Binding-dataset bias	High benchmark accuracy	Model may exploit dataset regularities rather than transferable interaction signal	Bias-aware or externally shifted evaluation
Inadequate distribution-shift definition	Stable performance under one convenient split	Generalization claim may not match deployment conditions	Explicitly specified OOD regimes
In-distribution robustness only	Strong internal ranking	Performance may deteriorate or reorder under chemical shift	Independent OOD comparison
Analogue or series leakage	Strong classification accuracy	Memorization can be mistaken for chemical understanding	Similarity-aware split and matched counterfactual tests
Stereo-aware input without stereo-specific ablation	Apparent benefit of a chiral model	Configuration may not be the feature responsible for improvement	Proposed: paired stereo-aware/stereo-blinded comparison

Note: The final row is a proposed validation requirement derived from the manuscript; it is not presented as an established benchmark standard.

Figure 2 integrates the completed data-integrity and evaluation argument as a braided validation pathway rather than a single benchmark pipeline.

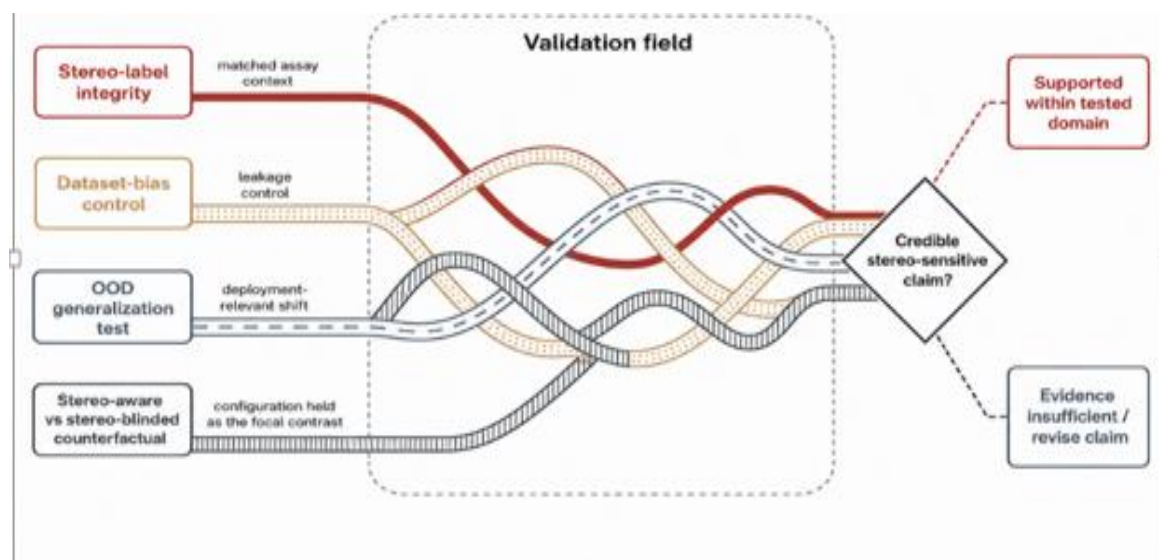


Figure 2. Braided validation pathway for a credible stereo-sensitive prediction claim

A framework for explicit stereochemical representation

The preceding analysis supports a proposed framework in which stereochemistry is treated as a controlled information variable across the machine-learning lifecycle. Natural-product AI already faces domain-specific problems of chemical diversity, data sparsity, characterization, and heterogeneous evidence [27]. The framework therefore does not prescribe a single architecture. It specifies conditions that a workflow should satisfy before claiming to model stereochemical dependence.

First, the dataset should declare the stereochemical state represented by every model input: experimentally assigned, unspecified, mixed, interconverting, or computationally enumerated. Second, that state should remain distinguishable through structure resolution, canonicalization, featurization, augmentation, and model ingestion. This matters because different representations

induce different effective neighborhood structures and therefore change the learning problem rather than merely changing file format [28].

Third, representation depth should follow the endpoint. A stereo-preserving graph may be sufficient when configurational identity matters but explicit conformational geometry does not demonstrably add information. Geometry-aware or conformer-aware views can be added when the mechanism or benchmark justifies them. Multimodal molecular learning demonstrates that complementary representations can be integrated rather than forcing all chemical information into one carrier [29]. In this article, such integration is an option, not an assumed advantage.

Fourth, learned invariances must be chemically audited. Contrastive molecular learning depends on augmentation choices that define which transformations the model is encouraged to ignore [30]. The framework therefore

proposes that stereo-altering augmentations be prohibited or explicitly tested whenever the label may depend on configuration.

Finally, model evaluation should ask whether stereochemical information changes predictions under controlled counterfactuals and relevant domain shifts. The framework thus defines explicit stereo representation as a precondition for testing stereochemical dependence, not as proof that such dependence exists. A negative stereo-specific result can itself be scientifically informative if the representation, labels, and validation design were capable of detecting a difference.

Boundary conditions and computational trade-offs

Greater structural detail has computational and epistemic costs. Conformer generation exemplifies this trade-off: different methods and ensemble sizes vary in structural coverage and quality, and generating additional conformers does not guarantee that biologically relevant geometries are represented [31]. For flexible natural products, exhaustive treatment may be expensive or unrealistic. A validated stereo-preserving two-dimensional model can therefore be preferable to an inadequately sampled three-dimensional model when the latter adds uncertain coordinates rather than reliable endpoint information.

The need for geometry is also endpoint-specific. Geometry-enhanced graph learning has been applied successfully to chromatographic enantioseparation, an analytical problem in which stereoselective interactions directly shape retention [32]. That result supports explicit geometry for that task but cannot be transferred automatically to receptor binding, functional activity, or selectivity, because chromatographic stationary phases and experimental conditions define a different interaction system.

Recent CNS-penetration modeling likewise illustrates that stereo-aware methods can be designed for developability endpoints and broader chemical space [33]. CNS penetration, however, integrates physicochemical and biological processes different from target affinity. Its inclusion therefore strengthens the boundary argument: stereochemistry may be relevant across multiple prediction domains, but the representation and validation requirements must follow the endpoint actually being modeled.

The practical rule is consequently conditional. Preserve configuration whenever it could matter; add geometry when the endpoint justifies geometric information; add conformer ensembles when conformational multiplicity is itself relevant; and reject additional complexity when it cannot demonstrate incremental, out-of-distribution value.

Limitations

The proposed framework is conceptual and has not been prospectively validated as a unified workflow for complex natural products. Its evidence comes from heterogeneous computational, analytical, medicinal-chemistry, database, and methodological settings. Transfer from an enantioselective chromatographic task, an atropisomeric kinase series, or a general molecular benchmark to another natural-product class or biological endpoint therefore requires direct testing.

Uncertainty also remains incompletely represented by point predictions. Chemical machine-learning uncertainty depends on data, model, and distribution shift, and calibration behavior can vary substantially across settings [34]. Stereo-specific confidence should consequently not be inferred from ordinary predictive confidence alone. Comparative work on neural uncertainty quantification further shows that no single uncertainty method is uniformly reliable across molecular-property tasks [35]. Several difficult cases remain outside any simple encoding rule: uncertain absolute configuration, dynamic stereochemical interconversion, protonation and tautomerism, multiple conformers, protein flexibility, assay disagreement, sparse matched stereoisomer pairs, and natural-sample mixtures. Explicit stereochemistry can prevent one class of information loss, but it cannot resolve these uncertainties automatically. The framework therefore requires empirical falsification rather than acceptance as a default performance guarantee.

CONCLUSION

Stereochemistry should be treated as model information before it is treated as a model feature. When binding, bioactivity, selectivity, or another endpoint can depend on configuration, collapsing stereoisomers upstream makes stereo-sensitive inference impossible to test cleanly. Explicit representation preserves that scientific question; it does not predetermine the answer.

For complex natural products, this principle extends from structure curation and provenance through graph or geometric representation to validation against matched counterfactuals and distribution shift. Stereo-preserving two-dimensional models, geometry-aware networks, and conformer ensembles are complementary tools whose value depends on the endpoint, data quality, and available structural evidence. The proposed framework therefore rejects both stereo blindness and indiscriminate three-dimensional complexity.

Its strongest claim is methodological: models should not be credited with learning stereochemical pharmacology unless their data, representations, and evaluations make stereochemical distinctions identifiable. Whether those

distinctions improve prediction, explain selectivity, or generalize prospectively remains an empirical question that must be tested rather than assumed.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] McGibbon M, Shave S, Dong J, Gao Y, Houston DR, Xie J, et al. From intuition to AI: Evolution of small molecule representations in drug discovery. *Brief Bioinform.* 2024;25(1). doi:10.1093/bib/bbad422
- [2] Boiko DA, Reschützeggger T, Sanchez-Lengeling B, Blau SM, Gomes G. Advancing molecular machine learning representations with stereoelectronics-infused molecular graphs. *Nat Mach Intell.* 2025;7(5):771-81. doi:10.1038/s42256-025-01031-9
- [3] Yan J, Wang H, Yang W, Ma X, Sun Y, Hu W. ChiGNN: Interpretable algorithm framework of molecular chiral knowledge-embedding and stereosensitive property prediction. *J Chem Inf Model.* 2025;65(7):3239-47. doi:10.1021/acs.jcim.4c02259
- [4] Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT. Natural products in drug discovery: Advances and opportunities. *Nat Rev Drug Discov.* 2021;20(3):200-16. doi:10.1038/s41573-020-00114-z
- [5] Bitchagno GTM, Nchiozem-Ngnitedem VA, Melchert D, Fobofou SA. Demystifying racemic natural products in the homochiral world. *Nat Rev Chem.* 2022;6(11):806-22. doi:10.1038/s41570-022-00431-4
- [6] Calcaterra A, Cianfoni G, Tortora C, Manetto S, Grassi G, Botta B, et al. Natural cannabichromene (CBC) shows distinct scalemicity grades and enantiomeric dominance in Cannabis sativa strains. *J Nat Prod.* 2023;86(4):909-14. doi:10.1021/acs.jnatprod.2c01139
- [7] Dadiotis E, Cui M, Gerasi M, Mitsis V, Melliou E, Makriyannis A, et al. A simple chiral 1H NMR method for the discrimination of (R)- and (S)-cannabichromene in complex natural mixtures and their effects on TRPA1 activity. *J Nat Prod.* 2024;87(1):77-84. doi:10.1021/acs.jnatprod.3c00796
- [8] Axen SD, Huang XP, Cáceres EL, Gendele L, Roth BL, Keiser MJ. A simple representation of three-dimensional molecular structure. *J Med Chem.* 2017;60(17):7393-409. doi:10.1021/acs.jmedchem.7b00696
- [9] Heid E, Greenman KP, Chung Y, Li SC, Graff DE, Vermeire FH, et al. Chemprop: A machine learning package for chemical property prediction. *J Chem Inf Model.* 2024;64(1):9-17. doi:10.1021/acs.jcim.3c01250
- [10] Coley CW, Green WH, Jensen KF. RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *J Chem Inf Model.* 2019;59(6):2529-37. doi:10.1021/acs.jcim.9b00286
- [11] Boldini D, Ballabio D, Consonni V, Todeschini R, Grisoni F, Sieber SA. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. *J Cheminform.* 2024;16(1):35. doi:10.1186/s13321-024-00830-3
- [12] Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, et al. Geometry-enhanced molecular representation learning for property prediction. *Nat Mach Intell.* 2022;4(2):127-34. doi:10.1038/s42256-021-00438-4
- [13] Hamakawa Y, Miyao T. Understanding conformation importance in data-driven property prediction models. *J Chem Inf Model.* 2025;65(7):3388-404. doi:10.1021/acs.jcim.5c00018
- [14] Morehead A, Cheng J. Geometry-complete perceptron networks for 3D molecular graphs. *Bioinformatics.* 2024;40(2). doi:10.1093/bioinformatics/btae087
- [15] Lu W, Zhang J, Huang W, Zhang Z, Jia X, Wang Z, et al. DynamicBind: Predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nat Commun.* 2024;15(1):1071. doi:10.1038/s41467-024-45461-2
- [16] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics.* 2018;34(21):3666-74. doi:10.1093/bioinformatics/bty374
- [17] McVicker RU, O'Boyle NM. Chirality of new drug approvals (2013-2022): Trends and perspectives. *J Med Chem.* 2024;67(4):2305-20. doi:10.1021/acs.jmedchem.3c02239
- [18] Bewley CA, Sulikowski GA, Yang ZJ, Bifulco G, Cho HM, Fullenkamp CR. Properties of configurationally stable atropenantiomers in drug discovery. *Acc Chem Res.* 2023;56(4):414-24. doi:10.1021/acs.accounts.2c00648
- [19] Chandrasekhar J, Dick R, Van Veldhuizen J, Koditek D, Lepist EI, McGrath ME, et al. Atropisomerism by design: Discovery of a selective and stable phosphoinositide 3-kinase (PI3K) β inhibitor. *J Med Chem.* 2018;61(15):6858-68. doi:10.1021/acs.jmedchem.8b00797
- [20] Chandrasekhar V, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, et al. COCONUT 2.0: A

- comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Res.* 2025;53(D1). doi:10.1093/nar/gkae1063
- [21] 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
- [22] Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, et al. The LOTUS initiative for open knowledge management in natural products research. *Elife.* 2022;11. doi:10.7554/eLife.70780
- [23] Graber D, Stockinger P, Meyer F, Mishra S, Horn C, Buller R. Resolving data bias improves generalization in binding affinity prediction. *Nat Mach Intell.* 2025;7(10):1713-25. doi:10.1038/s42256-025-01124-5
- [24] Tossou P, Wognum C, Craig M, Mary H, Noutahi E. Real-world molecular out-of-distribution: Specification and investigation. *J Chem Inf Model.* 2024;64(3):697-711. doi:10.1021/acs.jcim.3c01774
- [25] Fooladi H, Vu TNL, Mathea M, Kirchmair J. Evaluating machine learning models for molecular property prediction: Performance and robustness on out-of-distribution data. *J Chem Inf Model.* 2025;65(19):9871-91. doi:10.1021/acs.jcim.5c00475
- [26] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. *J Chem Inf Model.* 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403
- [27] Saldivar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F. Natural product drug discovery in the artificial intelligence era. *Chem Sci.* 2022;13(6):1526-46. doi:10.1039/D1SC04471K
- [28] Rottach F, Schieferdecker S, Eickhoff C. The topology of molecular representations and its influence on machine learning performance. *J Cheminform.* 2025;17(1):109. doi:10.1186/s13321-025-01045-w
- [29] Shen A, Yuan M, Ma Y, Du J, Wang M. Complementary multi-modality molecular self-supervised learning via non-overlapping masking for property prediction. *Brief Bioinform.* 2024;25(4). doi:10.1093/bib/bbae256
- [30] Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. *Nat Mach Intell.* 2022;4(3):279-87. doi:10.1038/s42256-022-00447-x
- [31] McNutt AT, Bisiriyu F, Song S, Vyas A, Hutchison GR, Koes DR. Conformer generation for structure-based drug design: How many and how good? *J Chem Inf Model.* 2023;63(21):6598-607. doi:10.1021/acs.jcim.3c01245
- [32] Xu H, Lin J, Zhang D, Mo F. Retention time prediction for chromatographic enantioseparation by quantile geometry-enhanced graph neural network. *Nat Commun.* 2023;14(1):3095. doi:10.1038/s41467-023-38853-3
- [33] Collins JW, Abedi M, Ramirez D, Philippe-Venec L, Moore B, Asher JO. CANDID-CNS: AI unlocks stereochemistry and beyond Rule of 5 to predict CNS penetration of small molecules. *J Chem Inf Model.* 2025;65(23):12918-28. doi:10.1021/acs.jcim.5c01888
- [34] Heid E, McGill CJ, Vermeire FH, Green WH. Characterizing uncertainty in machine learning for chemistry. *J Chem Inf Model.* 2023;63(13):4012-29. doi:10.1021/acs.jcim.3c00373
- [35] 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