Pharmacophore

ISSN-2229-5402

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

When Molecular Representations Conceal Pharmacological Differences: An Epistemological Analysis of Chemical Similarity, Biological Context, and Generalization Claims in Artificial Intelligence–Enabled Cheminformatics

Laura Mendez[1]*, Hiroshi Sato[2], Anna Kowalczyk[3], David Chen[4]

1. ***Department of Cheminformatics and Molecular Representation, IE Business School, Madrid, Spain.***
2. ***Department of Computational Biology and Drug Informatics, Keio University, Tokyo, Japan.***
3. ***Department of Pharmaceutical Informatics and Molecular Modeling, University of Warsaw, Warsaw, Poland.***
4. ***Department of Chemical Informatics and AI in Drug Discovery, National University of Singapore, Singapore.***

ARTICLE INFO

Received:
10 May 2026
Received in revised form:
07 September 2026
Accepted:
09 September 2026
Available online:
28 October 2026

Keywords: Molecular representation, Chemical similarity, Pharmacological similarity, Biological context, Activity cliffs, Out-of-distribution generalization

ABSTRACT

Artificial intelligence–enabled cheminformatics depends on transforming molecules into representations that models can compare and generalize from. Yet similarity induced by a representation is not equivalent to pharmacological similarity. This conceptual analysis examines how molecular encodings preserve, compress, or omit information; how biological context changes the meaning of similarity; and how these choices delimit defensible generalization claims. It develops a representation–context–claim account in which inferential strength depends on whether the representation retains information relevant to the property of interest, whether the biological context defining similarity matches the intended claim, and whether validation reproduces the transfer boundary implied by that claim. Activity cliffs, phenotypic profiling, heterogeneous bioactivity records, and out-of-distribution evaluation illustrate different ways in which representational proximity can overstate pharmacological continuity. The resulting position is neither that chemical similarity is uninformative nor that increasingly expressive representations are intrinsically superior. Similarity becomes epistemically useful when its representational basis, biological reference frame, and intended inference are aligned. Generalization in cheminformatics should therefore be treated as a calibrated claim about preserved information under specified contextual and distributional conditions.

*This is an **open-access** article distributed under the terms of the Creative Commons Attribution-Non Commercial-Share Alike 4.0 License, which allows others to remix, and build upon the work non commercially.*

To Cite This Article: Mendez L, Sato H, Kowalczyk A, Chen D. When Molecular Representations Conceal Pharmacological Differences: An Epistemological Analysis of Chemical Similarity, Biological Context, and Generalization Claims in Artificial Intelligence–Enabled Cheminformatics. Pharmacophore. 2026;17(5):46-53. https://doi.org/10.51847/bpQbSP97Je

Introduction

Modern cheminformatics operates on representations rather than chemically complete objects. Strings, fingerprints, descriptors, graphs, images, geometric encodings, and learned embeddings make selected aspects of a molecule computationally accessible while leaving others compressed or absent. Reviews of molecular representation in drug discovery emphasize that these encodings preserve different chemical information and impose different inductive assumptions [1]. The broader representation literature likewise shows that no encoding is neutral to the learning problem; each establishes a particular computational view of molecular identity and similarity [2].

This matters because contemporary AI-enabled drug design is increasingly evaluated through the apparent power of its representations. Recent systematic evidence shows substantial heterogeneity in molecular encodings, architectures, generative approaches, and evaluation practices [3]. It is therefore tempting to treat newer or more highly pretrained representations as progressively closer approximations to pharmacologically meaningful latent structure. Direct comparative analysis challenges

Corresponding Author: Laura Mendez; Department of Cheminformatics and Molecular Representation, IE Business School, Madrid, Spain. E-mail: laura.mendez@ie.edu.

that assumption: pretrained molecular models do not invariably produce representations better aligned with downstream activity relationships than strong conventional encodings, and their advantages vary across targets and tasks [4]. Technical sophistication is therefore not equivalent to pharmacological adequacy.

The problem becomes sharper when models are asked to generalize. "Out of distribution" is not one condition but a family of shifts created by different partitioning and transfer assumptions. Comparative work shows that apparent robustness changes markedly across random, scaffold-based, cluster-based, and other distributional separations [5]. A claim that a representation generalizes is consequently incomplete unless the relevant shift is specified.

The central problem here is epistemological rather than merely architectural: if molecules are rendered similar by a representation, what pharmacological similarity, if any, may legitimately be inferred? Answering requires separating what the representation preserves, what biological context gives similarity its meaning, and what strength of claim is warranted by validation. The analysis first examines information retained or suppressed by major representation families, then traces where chemical and pharmacological similarity diverge across biological contexts. These foundations support a representation–context–claim account developed in the next section of the manuscript.

Molecular Representation and Pharmacological Meaning

A molecular representation is an information-bearing abstraction whose usefulness depends on what the task requires. This becomes visible when graph-centric assumptions are relaxed. Set representation learning has shown that useful prediction can sometimes be achieved from collections of atomic or bond-level invariants without making the full explicit graph the sole organizing structure [6]. Image-based approaches instead learn from rendered two-dimensional depictions and therefore expose regularities different from fingerprints or message-passing graphs [7]. These are not interchangeable descriptions of the same computational object: they create different neighborhoods of similarity.

Representations that extend beyond static two-dimensional structure make the same point. Molecular-video foundation models can encode sequences of conformational or geometric states unavailable to a single static graph [8]. Hierarchical multi-channel learning can also make structural representation responsive to property context rather than assuming one invariant encoding is optimal for every endpoint [9]. Yet richer representation remains selective. Dynamic encodings sample particular states and scales; hierarchical encodings privilege particular structural levels. Added information may increase relevance without guaranteeing that the information added is the information required by the pharmacological claim.

Conformation-aware pretraining further illustrates the distinction. Combining fingerprints, functional groups, topological distances, and three-dimensional geometry can improve selected property and activity-cliff tasks [10]. However, comparative work on learned and fixed representations shows that their relative utility remains dataset- and endpoint-dependent [11]. A learned embedding is therefore an empirical organization of training information under a specific objective, not a chemically complete state description.

Representations may also be enriched by interaction structure or chemical knowledge. PotentialNet demonstrated that graph-based interaction modeling can support molecular and protein–ligand property prediction [12], while knowledge-guided pretraining showed that structural and semantic chemical knowledge can improve selected downstream tasks [13]. Such enrichment expands what a model can express, but the evidential force of the resulting similarity remains bounded by the endpoint and validation regime in which it was established.

The resulting distinction is better expressed as retained, compressed, and omitted information than as a ladder from simple to advanced. **Table 1** maps major representation families to the information they privilege and the pharmacological inferences that become vulnerable when relevant information is absent.

Table 1. Information preservation and inferential blind spots across molecular representation families

Representation family	Information preferentially retained	Information compressed or indirect	Defensible similarity claim	Characteristic blind spot
Descriptors/fingerprints	Selected properties, fragments, local topology	Global geometry, dynamics, biological context	Proximity in the encoded feature space	Extending feature proximity beyond its defined information
Strings/images	Notation- or depiction-dependent structural patterns	Full 3D state and biological environment	Regularities learned from the chosen encoding	Treating notation or rendering conventions as invariant chemistry
Graph/set encodings	Connectivity, neighborhoods, composition or atomic invariants	Geometry and context unless supplied	Structure–property relations supported by represented chemistry	Assuming structural proximity entails similar biological action
3D/conformation-aware encodings	Coordinates, distances, angles, stereochemical states	Full conformational ensemble, solvent and target-induced states	Similarity in geometrically relevant states	Treating sampled geometry as the biologically realized state
Dynamic/multi-state encodings	Variation across represented	Unobserved states and biological response	Relations dependent on modeled structural variability	Equating modeled dynamics with

	conformations or states			pharmacodynamic behavior
Hierarchical/knowledge-guided encodings	Multi-scale structural or semantic relations	Information absent from hierarchy or knowledge source	Task-conditioned similarity supported by supplied structure/knowledge	Mistaking richer latent organization for unrestricted transferability

A representation may thus be sufficient for one property while suppressing distinctions essential to another. An encoding useful for solubility can be inadequate for target selectivity; a geometry-sensitive representation can still miss cellular-state dependence. Pharmacological meaning therefore arises from the relation between encoded information and the property for which similarity is claimed, not from representation complexity alone.

Similarity, Biological Context and Generalization

Activity cliffs provide a direct counterexample to equating chemical proximity with pharmacological continuity. They occur when structurally similar compounds show unexpectedly large differences in measured activity. Comparative analyses demonstrate that modern machine-learning approaches still struggle with these local structure–activity discontinuities [14]. Recent synthesis therefore treats activity cliffs as informative boundaries that expose where smooth similarity assumptions fail rather than as mere nuisance outliers [15]. A coherent neighborhood in representational space can contain a discontinuity in the biological property that motivated the comparison.

Learning objectives can be made sensitive to this problem. Activity-cliff-aware reinforcement learning can explicitly incorporate local discontinuities during molecular generation [16], and transfer learning has used activity-cliff prediction to inform drug–target interaction modeling [17]. These methods show that known representational failure modes can be made visible to the objective. They do not establish that cliff-aware neighborhoods generalize across targets, assays, cell states, toxicity endpoints, or therapeutic contexts.

Biological context introduces additional distinctions that structure alone cannot guarantee. Cell Painting has revealed biologically interpretable morphological responses among compounds whose prior assay histories provided limited evidence of activity [18]. In a large comparison involving 16,170 compounds and 270 assays, chemical structure, cellular morphology, and gene-expression profiles contributed partly complementary predictive information [19]. Chemical similarity concerns relations among molecular descriptions; pharmacological similarity concerns relations among biological effects defined within particular systems. Their correspondence therefore requires evidence.

Even the label “bioactivity” is context-laden. ChEMBL integrates measurements across heterogeneous targets, assay formats, action types, endpoints, and historical periods [20]. The meaning of similarity in such data depends on what was measured, against which biological entity, and under which conditions. Work on multimodal foundation models in molecular cell biology extends this principle by emphasizing the importance of cellular, spatial, transcriptional, and other state-defining modalities when biological meaning crosses levels [21]. Additional modalities do not eliminate ambiguity, but they expose information that structural similarity alone may not contain.

These constraints also limit the portability of machine-learning success. Critical analyses of AI in drug discovery emphasize that chemical and biological data differ in provenance, sparsity, measurement error, sampling logic, and semantic stability [22]. A model can therefore be internally coherent yet pharmacologically overextended if representation-induced similarity is interpreted beyond the biological context in which labels were generated. The operative question is not simply whether two molecules are similar, but similar in what representation, under what biological observation system, and for which intended inference.

Table 2 separates these relations by the object being compared and the boundary that limits the resulting claim.

Table 2. Biological reference frames that change the meaning of molecular similarity

Similarity relation	What is compared	Context fixing its meaning	Inference that may be supported	Boundary to stronger inference
Encoded chemical similarity	Molecules in a fingerprint, graph, embedding or other representation	Encoding and training objective	Proximity in that represented chemical space	Omitted structural, geometric or contextual information
Local structure–activity similarity	Closely related analogues	Target, assay, endpoint and chemical series	Activity continuity within a validated local SAR region	Activity cliffs and measurement uncertainty
Target-specific pharmacological similarity	Compounds measured against a target	Target identity, assay system and action type	Comparable target-directed behavior in the tested system	Polypharmacology, assay dependence and alternate mechanisms
Phenotypic similarity	Cellular morphology or expression responses	Cell type, dose, exposure time and platform	Comparable system-level response under matched conditions	Distinct mechanisms can converge phenotypically

Multimodal biological similarity	Molecular, phenotypic or transcriptional profiles	Modalities and their alignment	Cross-level resemblance supported by integrated evidence	Missing/discordant modalities and unresolved causal direction
Cross-domain predictive similarity	Training and test compounds treated as transferable	Scaffold/cluster separation, task definition and validation	Generalization to the shift reproduced by evaluation	Unseen chemical, assay or biological shifts

Generalization is therefore not defined by distance in chemical space alone. A model may transfer across scaffolds while failing across assay platforms, or preserve target-level activity while failing to preserve cellular phenotype. Conversely, molecules distant in one structural representation may converge functionally in a particular biological condition. The strength of a generalization claim must match the level at which similarity was observed and the boundary across which that similarity was actually tested.

A Representation–Context–Claim Framework

A defensible generalization claim requires more than a representation that performs well on an internal benchmark. Out-of-distribution molecular prediction illustrates why. Methods designed to preserve task-relevant semantic information under distribution shift can improve performance across selected datasets, but residual failures remain sensitive to how the shifted domain is constructed [23]. Likewise, hierarchical task-aware representations can reduce interference among distinct molecular-property labels by permitting representations to respond differently to property semantics [24]. These findings support a first condition of the present analysis: the representation used for a claim must remain informative for the property whose transfer is being asserted. Representational adequacy is therefore relational rather than intrinsic. A representation may be suitable for one endpoint while being epistemically insufficient for another.

A second condition concerns whether evaluation reproduces the chemical-domain separation implied by the claim. Large-scale drug-target benchmarks have shown that validation designs controlling chemical-series relatedness can yield materially different estimates from less stringent partitions [25]. Internal predictive performance can therefore reflect continuity between training and test chemical neighborhoods rather than evidence for transfer beyond them. This does not make conventional validation meaningless; it means that the validation boundary must correspond to the scientific statement being made. Claims about interpolation within familiar series, scaffold transfer, and transfer into structurally remote chemical space require different evidential support.

Representational richness and transfer validity must also remain analytically separate. Multidimensional learning that integrates different structural views can improve drug–drug interaction prediction for its tested endpoint [26]. Yet incorporating more molecular information does not demonstrate that a model will retain validity when the chemical domain changes. Scaffold-shift evaluation provides a more direct test of one such boundary because the structural organization of the evaluation set is deliberately separated from the training domain [27]. Even that test remains specific: scaffold shift does not reproduce changes in assay platform, target state, cellular environment, population, or therapeutic use. A model can therefore pass one transfer test while remaining unevaluated for another.

These distinctions permit a representation–context–claim account of cheminformatics inference. Its central proposition is that claim strength should be constrained jointly by three conditions: representational adequacy, meaning that the encoding preserves information relevant to the property under discussion; contextual compatibility, meaning that the biological system defining the evidence corresponds sufficiently to the system invoked by the claim; and transfer support, meaning that validation reproduces the chemical or biological displacement over which generalization is asserted. Contemporary perspectives on AI-enabled drug design similarly caution against separating computational outputs from experimental knowledge and the decisions into which those outputs enter [28]. Broader analyses of machine learning across drug discovery likewise show that usefulness varies with task, data provenance, validation strategy, and stage of development [29]. The framework developed here makes that conditionality explicit at the level of the claim itself.

The key consequence is non-compensatory. Strong performance in one dimension does not automatically repair a failure in another. A highly expressive representation cannot compensate for a mismatched assay context when the claim concerns another biological state. Rich phenotypic evidence cannot establish chemical-domain transfer if structurally novel compounds were never tested. Conversely, stringent scaffold validation cannot establish a pharmacological mechanism that the representation and measured endpoint do not encode. The resulting logic is therefore not a weighted score in which enough strength can overcome a missing evidential condition. Each condition limits a different inferential transition.

This relationship is summarized in **Figure 1**. The figure treats molecular representation, biological reference frame, and claim strength as intersecting analytical fields rather than a sequential pipeline.

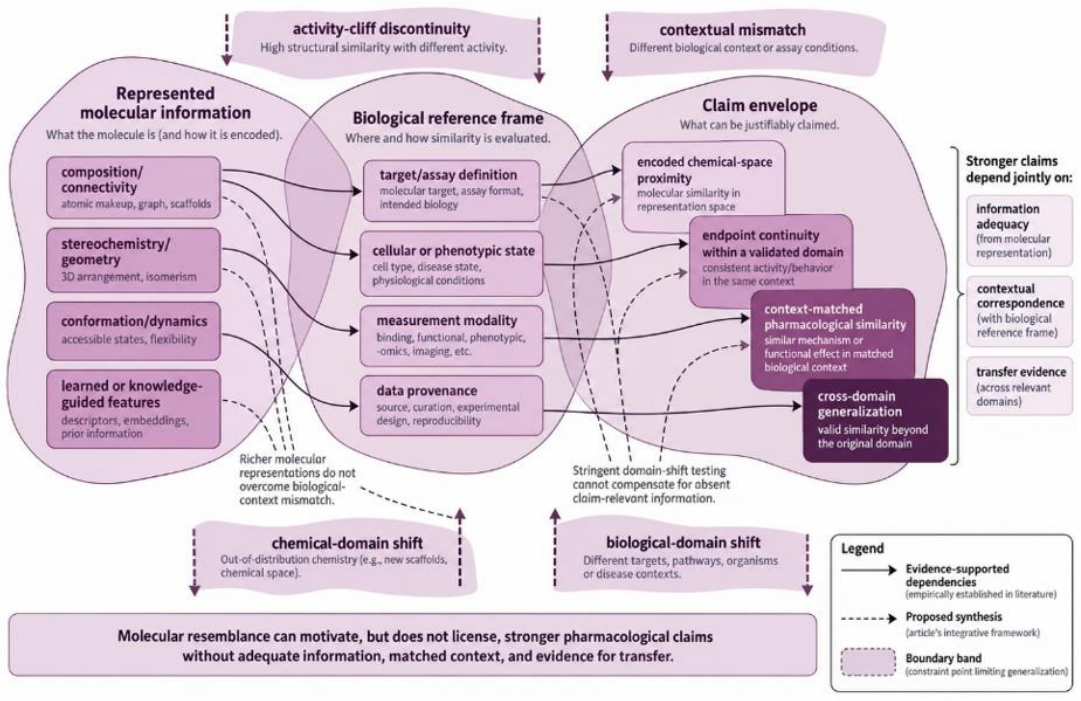


Figure 1. Where molecular resemblance stops licensing pharmacological generalization

Generalization failure can now be described more precisely as a mismatch among the three fields rather than as a unitary model defect. **Table 3** identifies recurrent configurations in which apparently persuasive predictive evidence supports a weaker claim than the one ultimately asserted.

Table 3. Generalization failures produced by misalignment among representation, biological context and claim strength

Failure configuration	Representation-side problem	Context violation	Overextended claim	Discriminating test or corrective response
Suppressed structural distinction	Claim-relevant stereochemical, geometric or dynamic information is absent or heavily compressed	Endpoint depends on the omitted distinction	Nearby encodings are interpreted as pharmacological equivalence	Re-evaluate with a representation that exposes the relevant structural variable
Activity-cliff smoothing	Local representation treats close analogues as smoothly related	Target/assay contains a sharp local SAR discontinuity	Structural proximity is taken to imply potency continuity	Conduct cliff-aware error analysis within the relevant series
Assay-semantic mismatch	Representation may be adequate for chemistry	Training labels arise from an assay different from the claimed biological state	Model output is generalized to a broader pharmacological effect	Test under the target assay or biological condition invoked by the claim
Phenotypic-context omission	Structure-only representation lacks system-response information	Cellular state materially alters observed effect	Chemical similarity is treated as equivalent biological response	Add context-bearing phenotypic or orthogonal biological evidence
Multimodal discordance	Modalities encode non-equivalent aspects of the system	Structure, phenotype or expression evidence disagree	Agreement in one modality is treated as system-wide similarity	Report concordance and discordance rather than collapsing modalities
Chemical-series leakage	Representation exploits near-neighbor continuity	Validation retains excessive series overlap	Internal accuracy is interpreted as chemical-domain generalization	Use cluster-, scaffold- or otherwise claim-matched separation
Scaffold-shift insufficiency	Representation survives one structural displacement	Biological or assay context also changes	Scaffold transfer is interpreted as universal OOD robustness	Test additional domain shifts corresponding to the intended use
Evidence-to-claim escalation	Representation and validation support a narrow endpoint	Claim moves to mechanism, clinical effect or broader pharmacology	Predictive association becomes stronger explanatory or translational inference	Restrict the claim or acquire evidence at the higher biological level

The table also clarifies why no universal numerical threshold can define representation–context–claim alignment. What counts as adequate information or relevant transfer depends on the scientific property and the inference at stake. The purpose of the framework is therefore not to certify models but to reveal when the evidence changes category: from encoded resemblance to

measured endpoint continuity, from endpoint continuity to pharmacological interpretation, or from within-domain prediction to genuine transfer.

Implications for Cheminformatics Research

The first implication concerns evaluation design. Out-of-distribution performance should be described by the actual displacement created, not by the generic label "OOD." The finding that different split strategies generate different estimates of robustness makes the tested boundary part of the scientific result rather than a secondary implementation detail [5]. Methods explicitly developed for semantic consistency under distribution shift reinforce this point because their performance remains defined relative to particular datasets and domain constructions [23]. For claims of scaffold transfer, chemical-series separation, target transfer or biological-context transfer, the evaluation should reproduce the corresponding displacement rather than substitute an easier one.

Validation should consequently be chosen after the intended claim is specified. Cluster-aware evaluation can reduce the optimism produced by chemically related training and test examples [25], while scaffold-shift testing directly interrogates a structural generalization boundary [27]. Neither is automatically superior in every project. Their value depends on whether they represent the chemical difference over which the claimed use requires inference. Reporting should therefore state both what was separated and what remained shared between training and evaluation domains.

A second implication concerns representation reporting. Model descriptions commonly identify the architecture while providing less explicit discussion of what information the molecular encoding can preserve. Yet comparative evidence shows that pretrained representations may fail to outperform conventional encodings on particular tasks [4], and context-dependent hierarchical representations can change which structural information becomes useful for different properties [9]. Researchers should therefore make the representational hypothesis explicit: which chemical distinctions are expected to matter for the endpoint, whether the representation can encode those distinctions, and which potentially relevant features remain unavailable.

Local failure analysis is particularly important when average performance can conceal pharmacologically meaningful discontinuities. Activity cliffs demonstrate that neighborhoods considered chemically close may contain precisely the cases in which smooth prediction is least reliable [14]. Evaluating errors by local structure–activity regime can therefore provide information that a global performance statistic cannot. This is not an argument for treating every model as an activity-cliff model; it is an argument for checking whether the intended inference is vulnerable to known violations of smooth chemical similarity.

A third implication concerns biological escalation. When the intended claim moves beyond a structurally defined endpoint toward cellular or pharmacological behavior, evidence should become correspondingly context-bearing. Phenotypic profiling can expose biological responses not evident from structural information alone [18], while combined structural and phenotypic analyses demonstrate that modalities can carry complementary predictive signal [19]. Bioactivity databases also encode substantial heterogeneity in targets, assays and measurement conditions [20]. A model trained on such information should not be described as learning an undifferentiated pharmacological property when its labels represent a particular measurement context.

Multimodal data can strengthen biological interpretation when the modalities correspond to the claim, but they should not be treated as an automatic remedy for uncertainty. Foundation-model work in molecular cell biology emphasizes that cellular state and modality materially shape biological representation [21], while critical analyses of drug-discovery AI show that chemical and biological data possess domain-specific limitations in provenance and interpretability [22]. The relevant question is therefore whether a new modality supplies information required by the inference, not whether adding modalities increases representational complexity.

Finally, model outputs should be reported in language proportional to what has been demonstrated. Drug-design perspectives increasingly situate AI inside iterative experimental and chemical reasoning rather than treating computational scores as autonomous evidence [28]. Across drug discovery and development, machine-learning value likewise varies with data, task, validation and decision stage [29]. For cheminformatics, this implies a practical hierarchy of wording. "Similar in the chosen representation" is weaker than "similar for the measured endpoint"; "similar for the measured endpoint" is weaker than "pharmacologically similar in the relevant biological context"; and all three are weaker than a claim of generalization to a new chemical or biological domain unless that transfer has been directly evaluated.

This hierarchy does not reduce the usefulness of similarity-based reasoning. It improves its precision. Chemical similarity remains one of the most productive organizing principles in medicinal chemistry and cheminformatics, but its inferential reach depends on how the molecule was represented, what biological question defined the labels, and which changes the model was actually shown to tolerate. Representation–context–claim alignment therefore functions as a research discipline for matching model evidence to scientific language.

Conclusion

Molecular representations conceal pharmacological differences when the similarity they induce is treated as stronger evidence than the encoded information and biological context justify. The problem cannot be resolved simply by choosing a more complex representation. Fingerprints, graphs, images, geometric encodings, dynamic representations, learned embeddings and

multimodal systems each preserve particular relations while suppressing others, and none converts representational proximity into pharmacological equivalence by itself.

Chemical similarity becomes pharmacologically informative only through a specified reference frame. Activity cliffs show that close structural neighbors can diverge sharply in biological activity; phenotypic and multimodal evidence show that biological resemblance can depend on information unavailable to structure alone; and domain-shift evaluation shows that model generalization is inseparable from the displacement actually tested.

The defensible strength of a generalization claim should therefore be bounded by three conditions: whether the representation retains information relevant to the property, whether the biological context matches the meaning assigned to similarity, and whether validation reproduces the transfer implied by the claim. Failure of any one condition narrows what may legitimately be inferred. The central epistemological requirement for AI-enabled cheminformatics is thus not maximal representational complexity, but explicit alignment between what is encoded, where biological meaning is established, and how far the resulting evidence is asked to travel.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: A review and practical guide. J Cheminform. 2020;12(1):56. doi:10.1186/s13321-020-00460-5
2. Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(5). doi:10.1002/wcms.1603
3. Abbasi K, Razzaghi P, Gharizadeh A, Ghareyazi A, Dehnad A, Rabiee HR, et al. Computational drug design in the artificial intelligence era: A systematic review of molecular representations, generative architectures, and performance assessment. Pharmacol Rev. 2026;78(1):100095. doi:10.1016/j.pharmr.2025.100095
4. Zhang Z, Bian Y, Xie A, Han P, Zhou S. Can pretrained models really learn better molecular representations for AI-aided drug discovery? J Chem Inf Model. 2024;64(7):2921-30. doi:10.1021/acs.jcim.3c01707
5. 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
6. Boulougouri M, Vandergheynst P, Probst D. Molecular set representation learning. Nat Mach Intell. 2024;6:754-63. doi:10.1038/s42256-024-00856-0
7. Li Y, Liu B, Deng J, Guo Y, Du H. Image-based molecular representation learning for drug development: A survey. Brief Bioinform. 2024;25(4). doi:10.1093/bib/bbae294
8. Xiang H, Zeng L, Hou L, Li K, Fu Z, Qiu Y, et al. A molecular video-derived foundation model for scientific drug discovery. Nat Commun. 2024;15(1):9696. doi:10.1038/s41467-024-53742-z
9. Wan Y, Wu J, Hou T, Hsieh CY, Jia X. Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation. Nat Commun. 2025;16(1):413. doi:10.1038/s41467-024-55082-4
10. Qiao J, Jin J, Wang D, Teng S, Zhang J, Yang X, et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. Nat Commun. 2025;16(1):4382. doi:10.1038/s41467-025-59634-0
11. 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
12. Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, et al. PotentialNet for molecular property prediction. ACS Cent Sci. 2018;4(11):1520-30. doi:10.1021/acscentsci.8b00507
13. Li H, Zhang R, Min Y, Ma D, Zhao D, Zeng J. A knowledge-guided pre-training framework for improving molecular representation learning. Nat Commun. 2023;14(1):7568. doi:10.1038/s41467-023-43214-1
14. 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
15. Szostek T, Szulczyk D. From obstacle to design advantage: Activity cliff aware modeling for small-molecule drug discovery. Drug Discov Today. 2026;31(1):104589. doi:10.1016/j.drudis.2025.104589
16. Hu X, Liu G, Zhao Y, Zhang H. Activity cliff-aware reinforcement learning for de novo drug design. J Cheminform. 2025;17(1):54. doi:10.1186/s13321-025-01006-3
17. Ibragimova R, Iliadis D, Waegeman W. Enhancing drug-target interaction prediction through transfer learning from activity cliff prediction tasks. J Chem Inf Model. 2025;65(13):6558-67. doi:10.1021/acs.jcim.5c00484

18. Pahl A, Liu J, Patil S, Rezaei Adariani S, Schölermann B, Warmers J, et al. Illuminating dark chemical matter using the Cell Painting assay. J Med Chem. 2024;67(11):8862-76. doi:10.1021/acs.jmedchem.4c00160
19. Moshkov N, Becker T, Yang K, Horvath P, Dancik V, Wagner BK, et al. Predicting compound activity from phenotypic profiles and chemical structures. Nat Commun. 2023;14(1):1967. doi:10.1038/s41467-023-37570-1
20. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, et al. The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52(D1). doi:10.1093/nar/gkad1004
21. Cui H, Tejada-Lapuerta A, Brbić M, Saez-Rodriguez J, Cristea S, Goodarzi H, et al. Towards multimodal foundation models in molecular cell biology. Nature. 2025;640(8059):623-33. doi:10.1038/s41586-025-08710-y
22. Bender A, Cortés-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
23. Wen X, Liu H, Long W, Wei S, Zhu R. Consistent semantic representation learning for out-of-distribution molecular property prediction. Brief Bioinform. 2025;26(2). doi:10.1093/bib/bbaf147
24. Kang L, Zhou S, Fang S, Liu S. Adapting differential molecular representation with hierarchical prompts for multi-label property prediction. Brief Bioinform. 2024;25(5). doi:10.1093/bib/bbae438
25. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441-51. doi:10.1039/c8sc00148k
26. Chen C, Shi X, Nie J, Xu J, Wang L. A molecular representation learning model based on multidimensional joint and cross-learning for drug-drug interaction prediction. J Chem Inf Model. 2025;65(16):8889-900. doi:10.1021/acs.jcim.5c01171
27. Iyer K, Sabar N. M-JEPA: Predictive self-supervised learning for molecular graphs with scaffold-shift evaluation on Tox21. J Chem Inf Model. 2026;66(14):8008-21. doi:10.1021/acs.jcim.6c00828
28. 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
29. 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