



Natural Product Macrocycles Break Conventional Molecular Featurization Through Conformational Complexity, Stereochemical Dependence, and Nonclassical Structure–Activity Relationships

Wei Zhang^{1*}, Chen Hui², Michael Tan¹

¹Department of Macrocyclic Chemistry and Conformational Analysis, Faculty of Pharmacy, University of Melbourne, Melbourne, Australia.

²Department of Nonclassical SAR for Macrocyclic Natural Products, Faculty of Pharmacy, National University of Singapore, Singapore.

ABSTRACT

Macrocycles challenge conventional molecular representation because their relevant properties depend less on a single static structure than on coupled stereochemistry, ring topology, conformational ensembles, intramolecular interactions, and environmental adaptation. This computational perspective examines why standard fingerprints, molecular graphs, and single-conformer descriptors become incomplete for natural-product macrocycles and cyclic peptides. The analysis integrates evidence on conformational sampling, chameleonicity, stereochemical encoding, dynamic polarity, membrane permeability, three-dimensional learning, and macrocyclic structure–activity relationships, while separating established observations from proposed computational implications. The central argument is that representation should be conditioned on the property being predicted and on whether that property is sensitive to molecular state, environment, or long-range ring organization. A proposed macrocycle-specific framework treats constitutional identity, stereochemistry, and ring topology as invariant information, with conformational-ensemble, environment-responsive, and uncertainty information added when mechanistically justified. Representation adequacy should be tested through matched ablations and scaffold-aware validation rather than inferred from model complexity. Principal limitations include the peptide-dominant evidence base, heterogeneous permeability measurements, force-field uncertainty, sparse coverage of nonpeptidic macrocycles, and limited prospective comparisons. This perspective defines a testable computational agenda rather than a validated universal standard, with prospective comparisons needed before any representation strategy can be preferred across macrocycle classes.

Key Words: *Macrocycles, Natural products, Molecular representation, Conformational ensembles, Stereochemistry, Molecular machine learning*

eIJPPR 2025; 15(3):88-97

HOW TO CITE THIS ARTICLE: Zhang W, Hui C, Tan M. Natural Product Macrocycles Break Conventional Molecular Featurization Through Conformational Complexity, Stereochemical Dependence, and Nonclassical Structure–Activity Relationships. *Int J Pharm Phytopharmacol Res.* 2025;15(3):88-97. <https://doi.org/10.51847/nwPFfxro7g>

INTRODUCTION

Macrocycles are not simply larger small molecules bent into rings. Their restricted yet often multistate conformational landscapes, beyond-rule-of-five physicochemical profiles, and capacity to reorganize exposed polarity create structure–property problems that

are difficult to summarize with a single static molecular view. Contemporary medicinal-chemistry analysis identifies macrocycles as a chemically distinctive design space in which ring architecture and three-dimensional behavior materially influence developability [1]. Cyclic peptides sharpen this problem because constrained backbones can support high-affinity recognition while

Corresponding author: Wei Zhang

Address: Department of Macrocyclic Chemistry and Conformational Analysis, Faculty of Pharmacy, University of Melbourne, Melbourne, Australia..

E-mail: ✉ wei.zhang@unimelb.edu.au

Received: 01 February 2025; **Revised:** 23 May 2025; **Accepted:** 25 May 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.



permeability and oral exposure remain difficult, sequence- and structure-dependent objectives [2]. The computational consequence proposed here is narrower than claiming that conventional descriptors universally fail: macrocycles increase the probability that a representation omits property-relevant state information. This risk is greatest when the same covalent graph can support materially different exposed surfaces, intramolecular contacts, or conformational populations under the conditions relevant to measurement. Featurization must therefore be judged against the endpoint, not against structural richness in the abstract.

Macrocyclic peptides also demonstrate why this question cannot be reduced to permeability alone. Their utility has been associated with access to binding surfaces that are difficult for conventional small molecules, while optimization must simultaneously negotiate potency, selectivity, stability, and delivery [3]. The broader peptide-drug landscape likewise shows that advances in discovery and modality engineering do not remove fundamental delivery constraints [4]. These observations matter for featurization because a representation that succeeds for one endpoint may discard information needed for another. A local fingerprint could remain adequate for a property dominated by constitution, yet become insufficient when a small modification reorganizes the entire ring or changes which polar groups are solvent exposed. The challenge is thus not molecular size itself but the possibility that local chemical identity and global property-relevant state become weakly separable.

Natural-product cyclic peptides provide the most instructive boundary case. They demonstrate that orally relevant behavior can exist in chemical space that exceeds conventional small-molecule heuristics, but they do not establish a general rule that natural macrocycles are orally bioavailable [5]. Their importance for cheminformatics lies instead in the coexistence of structural constraint and adaptive behavior. This perspective therefore asks a representation-level question: when does a molecular encoding cease to be an adequate surrogate for the experimentally relevant state of a macrocycle? The proposed answer is property-conditioned. Constitutional, topological, stereochemical, conformational, environmental, and uncertainty information should be treated as separable representational layers whose necessity must be tested rather than assumed.

Conformational ensembles and chameleonicity

A macrocycle can possess a fixed covalent graph while presenting several accessible three-dimensional states. Explicit conformational sampling of peptidic macrocycles shows that identifying and thermodynamically characterizing those states is itself a nontrivial

computational problem [6]. For cyclosporin, conformational flexibility has been directly connected to membrane-permeability behavior, providing a scaffold-specific example in which accessible states are not merely geometric noise [7]. The implication is that a canonical conformer may be an incomplete unit of representation whenever the predicted property depends on population shifts, transient shielding, or access to alternative global arrangements. It does not follow that every endpoint requires a full ensemble.

Chameleonicity makes that distinction sharper because the relevant distribution can change with environment. Simulations of macrocycles have demonstrated redistribution of conformational populations across conditions in a manner consistent with chameleonic behavior [8]. In representation terms, the molecule cannot always be described adequately by asking only “what shape does it have?”; one may instead need to ask “which states are accessible under the environment relevant to this endpoint?” This is especially important for properties involving transfer between aqueous, membrane-like, or binding environments. However, solvent models, force fields, and sampling completeness can alter inferred populations, so environmental responsiveness should be encoded only with explicit uncertainty.

Ensemble-aware learning is also becoming technically feasible at useful scale. CREMP provides large conformer–rotamer ensembles for macrocyclic peptides rather than assigning one geometry to each structure [9]. Across permeability-focused experiments, simulation, and ensemble-resource development, conformational heterogeneity is a material variable rather than a nuisance [7-9]. The proposed computational distinction is therefore between invariant identity features and state-distribution features. Ensemble summaries, state-conditioned embeddings, or multiple-instance architectures may be justified when state sensitivity is plausible, but their value must be demonstrated against simpler representations because additional conformers can also import noise, bias, and substantial computational cost.

Stereochemical density and long-range structural dependence

Stereochemistry creates a different information-loss problem. A molecular string can contain stereochemical symbols without guaranteeing that a learning architecture will use them effectively. Transformer studies have shown difficulty in chirality recognition from molecular string representations under evaluated conditions [10]. This finding is not macrocycle-specific, but it establishes an important computational premise: formally present stereochemical information may remain functionally inaccessible to a model. For macrocycles, that risk is

amplified conceptually because stereocenters are embedded in a closed ring where local configurations can help determine global backbone organization rather than acting as isolated substituent labels.

Graph models do not remove this concern automatically. Systematic evaluation of atomic featurization has shown that graph-neural-network performance depends on which atom-level features are supplied, demonstrating that connectivity alone does not define a complete predictive representation [11]. In a macrocycle, the relevant question is not merely whether chirality is encoded at each atom, but whether the architecture can propagate its consequences across the ring and distinguish stereochemical patterns that reorganize distant torsions, intramolecular contacts, or exposed surface. That long-range macrocycle interpretation is proposed here; the generic graph benchmark itself does not establish it. Ring closure makes this a particularly important hypothesis to test because information can propagate through both bonded topology and three-dimensional proximity, while ordinary message passing need not represent those two routes equivalently. Recent methods show plausible routes toward richer stereochemical encoding. ChiGNN explicitly embeds chiral knowledge for stereosensitive prediction and reports benefit in its evaluated tasks [12], while stereoelectronics-infused molecular graphs demonstrate that chemically enriched graph information can improve molecular learning beyond plain connectivity in selected settings [13]. Neither study validates a natural-product-macrocycle solution. Together, however, they support a design principle: stereo awareness should be treated as an explicit representational requirement when stereoisomers can differ functionally, and macrocycle-specific evaluation should test whether local chirality labels are sufficient or whether they require coupling to global geometry or ensemble information.

Ring topology, intramolecular interactions, and permeability

Permeability exposes the interaction among topology, conformation, and intramolecular polarity. In chameleonic compounds beyond the rule of five, dynamically exposed polarity has been associated with permeability and

solubility in ways that static scalar descriptions do not fully capture [14]. This does not establish a universal dynamic-polarity threshold. It instead shows that the effective polar surface presented by a flexible, internally hydrogen-bonded molecule can depend on molecular state. For macrocycles, a representation based only on fixed hydrogen-bond donor counts, polar surface area, or local fragments may therefore omit whether those groups are exposed, internally satisfied, or reorganized under the relevant environment.

Scaffold-level studies make the coupling more concrete. In cyclosporin O derivatives, changes in conformation, intramolecular hydrogen bonding, chameleonicity, membrane permeability, and cyclophilin A binding were interrelated across the investigated series [15]. Mechanistic simulations of conformationally flexible cyclic peptides likewise indicate structural adaptation during entry into and passage across lipid membranes [16]. These findings do not imply one universal permeation route, but they show why ring topology is computationally consequential: a local chemical modification can alter the intramolecular interaction network and thereby redistribute global states. This local-to-global dependence is exactly the kind of relationship that additive fragment representations may struggle to expose directly.

A particularly difficult case is provided by a large macrocyclic peptide whose membrane permeability was associated with a saddle-shaped conformation [17]. The defensible inference is not that saddle geometry is a transferable permeability rule, but that whole-ring organization can become a dominant explanatory variable in particular chemical series. Macrocyclic permeability can shift with global conformation and intramolecular polarity organization rather than with simple scalar descriptors alone [14–16]. Taken together, the evidence motivates a method–use-case distinction: representations should be selected according to whether the endpoint is expected to depend mainly on invariant constitution, explicit stereochemistry, a particular three-dimensional state, or an environment-conditioned ensemble. **Table 1** summarizes that distinction without assigning universal superiority to any method.

Table 1. Representation methods matched to macrocycle structure–property use cases

Method	Manuscript-derived use case	Retained information	Main failure risk	Evidence anchor/status
Constitution/stereo-aware 2D encoding	Screening when state dependence is unproven	Connectivity and explicit local stereochemistry	Global ring consequences remain latent	Proposed baseline; stereo/feature sensitivity supported
Single 3D conformer	Rigid or state-selective structural questions	One global geometry	Wrong conformer or environment	Proposed conditional use; whole-shape relevance illustrated
Conformational ensemble	State-sensitive property questions	Multiple accessible geometries	Sampling bias and cost	Macrocycle relevance supported

Environment-conditioned ensemble	Permeability involving shielding, intramolecular hydrogen bonds, or membrane adaptation	State redistribution and effective polarity	Solvent, membrane, or assay mismatch	Evidence-supported premise
----------------------------------	---	---	--------------------------------------	----------------------------

Note: Method-to-use-case matching is an original proposed synthesis, not a validated decision rule. Evidence anchors support the stated premises, not comparative predictive superiority.

Limits of fingerprints, graphs, and single-conformer descriptors

Conventional representations fail differently rather than uniformly. Fingerprints efficiently encode local substructures and remain useful when the endpoint is dominated by constitution or when data are too sparse to justify high-dimensional geometric models. Molecular graphs add relational structure and can learn task-specific features, while strings provide compact sequence-like encodings. Yet each representation exposes some molecular distinctions more directly than others [18]. The macrocycle problem is therefore one of information sufficiency: a representation may identify the molecule correctly while still suppressing the state variables through which a property emerges.

Representation choice also imposes inductive bias, so apparent model superiority cannot be separated cleanly from what information the model receives [19]. Importantly, replacing 2D features with a single 3D conformer does not automatically solve this problem. Controlled property-prediction comparisons show that 3D descriptors can help or hurt depending on the task and the quality or relevance of the conformer supplied [20]. A single geometry can therefore introduce false precision when the experimental endpoint integrates multiple states. For macrocycles, the appropriate baseline is not “2D versus 3D” in isolation, but the least complex representation that preserves the distinctions required by the endpoint. This criterion also separates representational incompleteness from ordinary estimation error: a weak model may fail despite adequate features, whereas a sophisticated model cannot recover information that the input never distinguishes. Macrocycle benchmarking should test both possibilities.

Three-dimensional and ensemble-aware representations

Geometric learning offers a principled way to expose molecular shape, distances, and spatial relationships directly to predictive models. Recent geometric deep-learning work demonstrates that such architectures can exploit three-dimensional information effectively across evaluated molecular-property tasks [21]. This establishes methodological capability, not a macrocycle-specific advantage. For macrocycles, geometry becomes compelling only when there is reason to expect that global ring shape, spatial shielding, or nonlocal contacts carry information unavailable from constitution and stereochemistry alone. The relevant comparison should

therefore include stereo-aware 2D baselines rather than assuming that geometric complexity is intrinsically more chemically faithful.

Three-dimensional contrastive learning provides another route by using structural views during representation learning rather than relying exclusively on handcrafted geometric descriptors [22]. Such pretraining could be valuable for macrocycles because multiple conformers can be treated as related realizations of one molecular identity. However, an ensemble is not merely data augmentation: conformers can differ in thermodynamic plausibility, environmental relevance, and property contribution. Treating every generated geometry as equivalent may blur rather than recover the underlying state distribution.

Self-conformation-aware pretraining further shows that conformational information can be integrated into learned representations while retaining downstream interpretability in evaluated settings [23]. The perspective proposed here extends that principle cautiously: macrocycle models should distinguish invariant identity from state-specific information and, when justified, learn across an ensemble with explicit aggregation or weighting. Candidate approaches include population-weighted descriptors, conformer-set pooling, state-conditioned embeddings, or paired representations across environments. Their comparative value remains an empirical question. The decisive test is whether these added state variables improve scaffold-aware external prediction, calibration, or mechanistic consistency enough to justify their uncertainty and computational expense. Ensemble aggregation should also preserve molecular identity across conformers rather than allowing the model to treat each geometry as an independent compound. Architectures that pool sets, learn state weights, or compare environments provide distinct hypotheses about how an endpoint integrates conformational information, and those hypotheses should be benchmarked separately.

Nonclassical SAR in macrocyclic chemical series

Macrocyclic structure–activity relationships can violate the intuition that small local edits should produce proportionally local property changes. De novo development of orally bioavailable small cyclic peptides shows that favorable exposure can emerge from coordinated structural design rather than from optimizing one scalar feature in isolation [24]. Here, “nonclassical SAR” is proposed as an interpretive term for series in which local chemical changes reorganize global

conformation, intramolecular contacts, permeability, or another developability property so strongly that fragment-level similarity becomes a poor surrogate for functional similarity. The term does not imply that macrocycle SAR is intrinsically discontinuous or that conventional medicinal-chemistry reasoning is invalid.

A hit-to-clinical-candidate cyclic-peptide program targeting KRAS further illustrates the multidimensional nature of macrocycle optimization, where potency, permeability, and developability were negotiated across an evolving series [25]. Such programs are important computationally because the label attached to one molecular structure may be the outcome of several coupled constraints. A model trained on only one endpoint can therefore learn a useful association while remaining blind to the structural reorganization that made the broader optimization successful. This is a reason to preserve

mechanistically plausible state information, not evidence that a particular model architecture will prospectively outperform alternatives.

The strongest local-to-global example in the selected evidence is a depsipeptide/peptide macrocycle series inspired by ent-verticilide, in which backbone constitution was shown to drive passive permeability independently of side-chain variation within the studied system [26]. This finding should not be generalized into side-chain irrelevance. It demonstrates instead that the location of predictive information can shift across series: sometimes side-chain chemistry dominates, whereas in another series backbone organization may control the phenotype. **Table 2** converts that observation into a decision matrix for representation selection, and **Figure 1** depicts the proposed dependency structure that makes macrocycle SAR computationally unusual

Table 2. Decision matrix for representing local edits that can produce global macrocycle property changes

Decision situation	Representation priority	Benefit	Trade-off or warning	Status
Matched analogs with stable global state	Constitution plus explicit stereochemistry	Preserves simple SAR signal efficiently	May miss hidden state changes	Proposed
Backbone-edited series	Ring topology plus global geometry	Captures scaffold reorganization	Conformer choice can dominate	Supported premise
Multi-property optimization	Shared invariant encoding plus endpoint-specific state features	Separates common identity from property context	Requires more labels and careful task definition	Proposed from and
Permeability-sensitive series	Ensemble and environment-responsive descriptors when justified	Can expose shielding and structural adaptation	Sampling and assay mismatch may add noise	Proposed from preceding permeability evidence

Note: The decision logic is an original synthesis. The cited studies support the underlying series-specific observations, not universal representation rankings.

Figure 1 shows how the manuscript's proposed interpretation connects apparently local macrocycle edits

to property-level consequences without assuming a deterministic pathway.

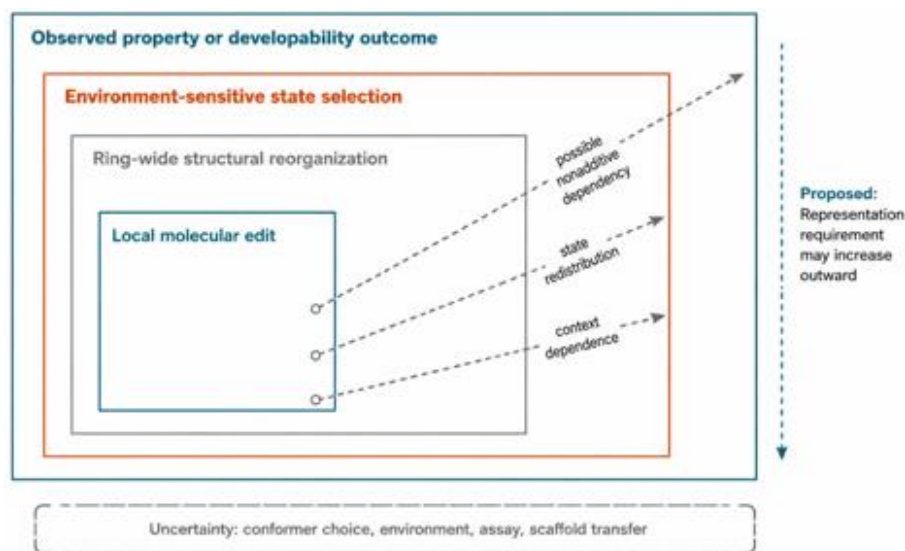


Figure 1. Local edits, global state reorganization, and nonclassical macrocycle SAR

A macrocycle-specific representation perspective

Macrocycle-specific machine learning already provides evidence that class-focused models can recover useful permeability-associated patterns. Work on membrane-permeating macrocycles derived design guidance from machine-learning analyses within the studied data domain [27]. The appropriate inference is bounded: dedicated models can exploit regularities in macrocycle data, but predictive associations need not identify the structural mechanism and may deteriorate outside the chemistry represented during training. A macrocycle-specific representation should therefore be defined by explicit hypotheses about information loss, not simply by training a larger model on a macrocycle-labelled subset.

Data infrastructure is equally important. CycPeptMPDB provides a dedicated resource for cyclic-peptide membrane-permeability modeling [28], but assembling measurements from different experimental contexts does not make those contexts chemically or analytically interchangeable. The proposed perspective therefore treats assay context and provenance as part of the prediction problem whenever heterogeneous labels are combined.

Model architecture cannot repair endpoint inconsistency that is hidden during dataset construction.

Multimodal cyclic-peptide learning demonstrates that complementary molecular views can be combined in one predictive framework [29]. Building on that feasibility, the principal proposal of this perspective is a layered, property-conditioned representation rather than a universal macrocycle fingerprint. Invariant constitution, explicit stereochemistry, and ring topology form the minimum molecular identity layer; conformational or ensemble information is added only when the endpoint is plausibly state-sensitive; environment-conditioned polarity or interaction features are added for context-dependent phenomena; and applicability-domain or uncertainty information accompanies predictions. **Table 3** organizes these layers as evidence-linked additions rather than mandatory complexity. The layers are deliberately gated. A model should advance to a more expensive layer only when the scientific endpoint makes the missing information plausible and an ablation shows that the addition contributes beyond simpler baselines. This prevents macrocycle-specific modeling from becoming synonymous with maximal featurization.

Table 3. Evidence-layer integration for the proposed property-conditioned macrocycle representation

Layer	Information represented	Add when	Main safeguard	Evidence status
Invariant identity	Constitution and ring connectivity	Always	Preserve a simple baseline	Established representation premise
Stereochemical/topological specificity	Chiral configuration and closed-ring organization	Stereoisomers or backbone organization may matter	Compare with explicit ablation	Proposed integration
Conformational state	Global geometry or conformer ensemble	Endpoint is plausibly state-sensitive	Test against simpler encodings	Supported feasibility
Environmental context	State-dependent polarity or interactions	Assay or membrane environment can change accessible states	Match representation to measurement context	Proposed integration
Domain and uncertainty	Chemical coverage and prediction support	Model is used beyond close training analogs	Report support separately from prediction	Proposed requirement

Note: The integrated layering is proposed and has not been prospectively validated as a unified architecture.

Benchmarking and generalization across rare scaffolds

Meaningful evaluation of the proposed perspective requires benchmark design that separates interpolation from chemically consequential extrapolation. MoleculeNet established the importance of explicit tasks, metrics, baselines, and data splits for molecular machine learning [30]. Comparative analysis of learned molecular representations likewise showed that representation performance varies with task and data regime rather than yielding one uniformly dominant encoding [31]. For macrocycles, these principles imply that stereo-aware 2D, single-conformer 3D, ensemble-aware, and multimodal representations should be compared under identical labels, splits, optimization budgets, and uncertainty reporting.

Direct evidence from cyclic-peptide permeability modeling makes scaffold separation especially important. A systematic benchmark of 13 AI methods found that generalization was materially harder under scaffold splitting than under random splitting in the evaluated data [32]. Benchmark conclusions are therefore sensitive to representation, split regime, and chemical-domain coverage [30–32]. Random splits remain useful for diagnosing interpolation, but they should not be allowed to stand in for rare-scaffold generalization. Macrocycle benchmarks should additionally prevent closely related stereoisomers or analog series from leaking across partitions when the scientific question concerns transfer to new structural families. Paired stereoisomers and matched

backbone edits deserve separate challenge subsets because they test whether the representation retains chemically small but structurally consequential distinctions. Reporting only global averages could conceal systematic failure exactly where macrocycle-specific representation is supposed to help.

A stronger prospective test would pair scaffold-aware splitting with external macrocycle collections and matched representation ablations. Success should be judged not only by average predictive error but by calibration, failure concentration, performance on rare ring systems, and the ability to preserve known stereochemical or conformational distinctions. If additional ensemble or environment-conditioned layers fail to improve these harder evaluations relative to simpler baselines, the proposed representation perspective should be narrowed rather than defended by architectural complexity. Conversely, improvement confined to random splits would be insufficient evidence for rare-scaffold transfer.

Boundary conditions and computational cost

State-aware representations carry a computational burden before model training begins. Macrocyclic conformational sampling across polar and apolar environments is sensitive to sampling strategy and can be difficult to converge [33]. Consequently, increasing the number of conformers is not equivalent to increasing useful information. Redundant, energetically implausible, or environment-inappropriate states can enlarge storage and training cost while making aggregation harder. A practical macrocycle workflow should therefore treat sampling depth as a scientific hyperparameter and justify it against the property being predicted. Cost should be measured across the complete representation pipeline, including conformer generation, solvent-specific simulation, feature calculation, storage, and repeated inference. This matters because a nominally accurate ensemble model may be unattractive if its marginal gain disappears after realistic sampling uncertainty is included.

Simulation quality also limits the credibility of ensemble-derived features. Validation of small-molecule force fields for macrocyclic compounds against NMR data in different solvents shows that agreement depends on the force field and environmental setting [34]. Experimental structural constraints can therefore be valuable when ensemble composition is central to the claim. They need not be available for every training molecule, but benchmark subsets with experimental ensemble evidence could distinguish failures of the learning architecture from failures in the conformers supplied to it.

Modern structure-prediction systems offer a different compromise. AlphaFold2-based approaches have been adapted for cyclic-peptide structure prediction and design

[35], expanding access to useful three-dimensional hypotheses. Yet a predicted structure is not synonymous with a thermodynamic ensemble or a solvent-dependent population. The proposed hierarchy is consequently economical: begin with constitution and stereochemistry, add predicted or generated 3D information when the endpoint warrants it, and reserve expensive environment-conditioned ensemble modeling for problems where simpler representations demonstrably leave systematic error. A practical stopping rule is therefore evidence-based rather than architectural: stop adding representational layers when external or scaffold-aware performance, calibration, or scientifically relevant error patterns no longer improve enough to offset uncertainty and cost.

Limitations

The first limitation is applicability domain. A dedicated methodology for cyclic-peptide membrane-permeability predictors shows why model outputs should be interpreted in relation to chemical support from the training domain [36]. Being classified as in-domain does not guarantee a correct prospective prediction, but ignoring domain altogether makes apparently precise outputs especially difficult to interpret for unusual natural-product scaffolds. Second, the evidence considered here is disproportionately peptide-centered. A recent permeability database for nonpeptidic macrocycles documents a distinct and heterogeneous chemical and assay space [37]. Principles derived from cyclic peptides may therefore transfer incompletely to polyketide, macrolide, depsipeptide, or other natural-product macrocycles. Permeability labels themselves can also differ by assay context, so prediction of one measurement should not be presented as a mechanistically equivalent prediction of in vivo exposure. Third, general molecular-learning studies continue to show that divergent tasks and distribution shift challenge generalization strategies [38]. This is a methodological boundary rather than direct macrocycle evidence. The perspective does not establish that its proposed layers will improve prospective prediction, nor that conformational complexity is the dominant source of error in every macrocycle dataset. Label noise, assay heterogeneity, model capacity, sparse stereochemical coverage, and chemical-class imbalance may equal or exceed representation error. Natural products also create practical representation challenges through unusual residues, noncanonical bonds, dense stereochemistry, and chemistry that may be poorly covered by generic preprocessing conventions. These limitations require explicit data audits before model failure is attributed to molecular featurization itself.

CONCLUSION

Natural-product macrocycles challenge conventional featurization not because two-dimensional chemistry becomes irrelevant, but because molecular identity can become insufficiently informative when a property depends on stereochemical organization, global ring geometry, conformational populations, or environmental adaptation. The central proposal is therefore conditional rather than architectural: represent invariant chemistry explicitly, then add state, environment, and uncertainty information only when the endpoint and evidence justify those layers.

This perspective also rejects a simple progression from fingerprints to graphs to three-dimensional models as though complexity guaranteed chemical fidelity. Single conformers can be misleading, ensembles can be poorly sampled, and richer representations can magnify data limitations. Credible macrocycle modeling consequently requires matched representation ablations, scaffold-aware evaluation, explicit applicability domains, and separation of structure prediction from ensemble thermodynamics.

The proposed framework remains a testable computational synthesis rather than a validated standard. Its strongest near-term use is to make representation assumptions visible and falsifiable. Prospective evaluation across chemically diverse natural-product macrocycles, multiple assay contexts, paired stereochemical or backbone perturbations, and external rare scaffolds is required before any layer can be considered generally necessary. A macrocycle representation should ultimately earn its complexity by improving reliable generalization and chemically meaningful discrimination, not by encoding more molecular detail for its own sake.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] Garcia Jimenez D, Poongavanam V, Kihlberg J. Macrocycles in drug discovery-learning from the past for the future. *J Med Chem*. 2023;66(8):5377-96. doi:10.1021/acs.jmedchem.3c00134
- [2] Ji X, Nielsen AL, Heinis C. Cyclic peptides for drug development. *Angew Chem Int Ed Engl*. 2024;63(3). doi:10.1002/anie.202308251
- [3] Vinogradov AA, Yin Y, Suga H. Macrocyclic peptides as drug candidates: Recent progress and remaining challenges. *J Am Chem Soc*. 2019;141(10):4167-81. doi:10.1021/jacs.8b13178
- [4] Muttenthaler M, King GF, Adams DJ, Alewood PF. Trends in peptide drug discovery. *Nat Rev Drug Discov*. 2021;20(4):309-25. doi:10.1038/s41573-020-00135-8
- [5] Naylor MR, Bockus AT, Blanco MJ, Lokey RS. Cyclic peptide natural products chart the frontier of oral bioavailability in the pursuit of undruggable targets. *Curr Opin Chem Biol*. 2017;38:141-7. doi:10.1016/j.cbpa.2017.04.012
- [6] Kamenik AS, Lessel U, Fuchs JE, Fox T, Liedl KR. Peptidic macrocycles-conformational sampling and thermodynamic characterization. *J Chem Inf Model*. 2018;58(5):982-92. doi:10.1021/acs.jcim.8b00097
- [7] Wang CK, Swedberg JE, Harvey PJ, Kaas Q, Craik DJ. Conformational flexibility is a determinant of permeability for cyclosporin. *J Phys Chem B*. 2018;122(8):2261-76. doi:10.1021/acs.jpcb.7b12419
- [8] Sethio D, Poongavanam V, Xiong R, Tyagi M, Vo DD, Lindh R, et al. Simulation reveals the chameleonic behavior of macrocycles. *J Chem Inf Model*. 2023;63(1):138-46. doi:10.1021/acs.jcim.2c01093
- [9] Grambow CA, Weir H, Cunningham CN, Biancalani T, Chuang KV. CREMP: Conformer-rotamer ensembles of macrocyclic peptides for machine learning. *Sci Data*. 2024;11(1):859. doi:10.1038/s41597-024-03698-y
- [10] Yoshikai Y, Mizuno T, Nemoto S, Kusuhara H. Difficulty in chirality recognition for Transformer architectures learning chemical structures from string representations. *Nat Commun*. 2024;15(1):1197. doi:10.1038/s41467-024-45102-8
- [11] Wojtuch A, Danel T, Podlowska S, Maziarka L. Extended study on atomic featurization in graph neural networks for molecular property prediction. *J Cheminform*. 2023;15(1):81. doi:10.1186/s13321-023-00751-7
- [12] 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
- [13] Boiko DA, Reschutze 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
- [14] Rossi Sebastiano M, Doak BC, Backlund M, Poongavanam V, Over B, Ermondi G, et al. Impact of dynamically exposed polarity on permeability and solubility of chameleonic drugs beyond the Rule of 5.

- J Med Chem. 2018;61(9):4189-202. doi:10.1021/acs.jmedchem.8b00347
- [15] Lee D, Lee S, Choi J, Song YK, Kim MJ, Shin DS, et al. Interplay among conformation, intramolecular hydrogen bonds, and chameleonicity in the membrane permeability and cyclophilin A binding of macrocyclic peptide cyclosporin O derivatives. *J Med Chem*. 2021;64(12):8272-86. doi:10.1021/acs.jmedchem.1c00211
- [16] Linker SM, Schellhaas C, Kamenik AS, Veldhuizen MM, Waibl F, Roth HJ, et al. Lessons for oral bioavailability: How conformationally flexible cyclic peptides enter and cross lipid membranes. *J Med Chem*. 2023;66(4):2773-88. doi:10.1021/acs.jmedchem.2c01837
- [17] Faris JH, Adaligil E, Popovych N, Ono S, Takahashi M, Nguyen H, et al. Membrane permeability in a large macrocyclic peptide driven by a saddle-shaped conformation. *J Am Chem Soc*. 2024;146(7):4582-91. doi:10.1021/jacs.3c10949
- [18] 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
- [19] 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
- [20] 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
- [21] Dobbelaere MR, Lengyel I, Stevens CV, Van Geem KM. Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. *J Cheminform*. 2024;16(1):99. doi:10.1186/s13321-024-00895-0
- [22] Moon K, Im HJ, Kwon S. 3D graph contrastive learning for molecular property prediction. *Bioinformatics*. 2023;39(6). doi:10.1093/bioinformatics/btad371
- [23] 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
- [24] Merz ML, Habeshian S, Li B, David JAGL, Nielsen AL, Ji X, et al. De novo development of small cyclic peptides that are orally bioavailable. *Nat Chem Biol*. 2024;20(5):624-33. doi:10.1038/s41589-023-01496-y
- [25] Tanada M, Tamiya M, Matsuo A, Chiyoda A, Takano K, Ito T, et al. Development of orally bioavailable peptides targeting an intracellular protein: From a hit to a clinical KRAS inhibitor. *J Am Chem Soc*. 2023;145(30):16610-20. doi:10.1021/jacs.3c03886
- [26] Thorpe MP, Smith AN, Blackwell DJ, Hopkins CR, Knollmann BC, Akers WS, et al. The backbone constitution drives passive permeability independent of side chains in depsipeptide and peptide macrocycles inspired by ent-verticillide. *Chem Sci*. 2024;15(36):14977-87. doi:10.1039/D4SC02758B
- [27] Williams-Noonan BJ, Speer MN, Le TC, Sadek MM, Thompson PE, Norton RS, et al. Membrane permeating macrocycles: Design guidelines from machine learning. *J Chem Inf Model*. 2022;62(19):4605-19. doi:10.1021/acs.jcim.2c00809
- [28] Li J, Yanagisawa K, Sugita M, Fujie T, Ohue M, Akiyama Y. CycPeptMPDB: A comprehensive database of membrane permeability of cyclic peptides. *J Chem Inf Model*. 2023;63(7):2240-50. doi:10.1021/acs.jcim.2c01573
- [29] Yu Y, Gu M, Guo H, Deng Y, Chen D, Wang J, et al. MuCoCP: A priori chemical knowledge-based multimodal contrastive learning pre-trained neural network for the prediction of cyclic peptide membrane penetration ability. *Bioinformatics*. 2024;40(8). doi:10.1093/bioinformatics/btae473
- [30] 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
- [31] Yang K, Swanson K, Jin W, Coley CW, 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
- [32] Liu W, Li J, Verma CS, Lee HK. Systematic benchmarking of 13 AI methods for predicting cyclic peptide membrane permeability. *J Cheminform*. 2025;17(1):129. doi:10.1186/s13321-025-01083-4
- [33] Tang X, Kokot J, Waibl F, Fernandez-Quintero ML, Kamenik AS, Liedl KR. Addressing challenges of macrocyclic conformational sampling in polar and apolar solvents: Lessons for chameleonicity. *J Chem Inf Model*. 2023;63(22):7107-23. doi:10.1021/acs.jcim.3c01123
- [34] Waibl F, Casagrande F, Dey F, Riniker S. Validating small-molecule force fields for macrocyclic compounds using NMR data in different solvents. *J Chem Inf Model*. 2024;64(20):7938-48. doi:10.1021/acs.jcim.4c01120
- [35] Rettie SA, Campbell KV, Bera AK, Kang A, Kozlov S, Bueso YF, et al. Cyclic peptide structure prediction and design using AlphaFold2. *Nat Commun*. 2025;16(1):4730. doi:10.1038/s41467-025-59940-7

- [36] Geylan G, De Maria L, Engkvist O, David F, Norinder U. A methodology to correctly assess the applicability domain of cell membrane permeability predictors for cyclic peptides. *Digit Discov.* 2024;3(9):1761-75. doi:10.1039/D4DD00056K
- [37] Feng Q, De Chavez D, Kihlberg J, Poongavanam V. A membrane permeability database for nonpeptidic macrocycles. *Sci Data.* 2025;12(1):10. doi:10.1038/s41597-024-04302-z
- [38] Garcia-Ortegon M, Seal S, Rasmussen C, Bender A, Bacallado S. Graph neural processes for molecules: An evaluation on docking scores and strategies to improve generalization. *J Cheminform.* 2024;16(1):115. doi:10.1186/s13321-024-00904-2