



Self-Supervision Before Screening for Learning Natural Product Representations from Unlabelled Molecular and Spectral Collections

James O'Leary^{1*}, Aisling Dunne¹, Sean O'Brien²

¹Department of Self-Supervised Learning for Natural Products, School of Pharmacy, Trinity College Dublin, Dublin, Ireland.

²Department of Molecular and Spectral Representation Learning, School of Pharmacy, University College Cork, Cork, Ireland.

ABSTRACT

Natural-product discovery faces a fundamental asymmetry: chemical information is abundant, yet biological labels are sparse, heterogeneous, and task-specific. Molecular structures, provenance data, three-dimensional conformations, and experimental spectra thus offer substantial reusable knowledge before any bioactivity endpoint is assigned. This article develops a computational framework that treats such unlabelled data as a pretraining substrate rather than as material awaiting conventional screening. The framework distinguishes learning signals from molecular graphs, sequences, geometries, and spectra, and it proposes chemically informed pretext objectives that preserve distinctions relevant for later pharmacological reasoning while avoiding artifacts from dataset biases. Experimental spectra are treated as a complementary chemical view because they encode composition and structure under measurement-specific conditions, and cross-modal learning is positioned to relate partially shared rather than interchangeable representations. The proposed structure separates pretraining, chemically meaningful representation testing, rare-label fine-tuning, and transfer evaluation across scaffold, target, and experimental domains. Chemical meaningfulness is not reduced to benchmark accuracy; representations are instead examined for structural sensitivity, coherent chemical neighbourhoods, modality correspondence, resistance to shortcuts, and behaviour outside the training distribution. This framework remains methodological rather than a validated universal architecture. Its practical utility depends on data provenance, spectral quality, conformational assumptions, label definitions, chemical-space coverage, and prospective testing on genuinely unseen natural products.

Key Words: Self-supervised learning, Natural products, Molecular representation, Spectroscopy, Chemical informatics, Multimodal learning

eIJPPR 2025; 15(5):82-90

HOW TO CITE THIS ARTICLE: O'Leary J, Dunne A, O'Brien S. Self-Supervision Before Screening for Learning Natural Product Representations from Unlabelled Molecular and Spectral Collections. Int J Pharm Phytopharmacol Res. 2025;15(5):82-90. <https://doi.org/10.51847/3s7itGuPN>

INTRODUCTION

Natural-product informatics increasingly begins with a paradox of abundance. Large open resources now organize substantial numbers of natural-product structures together with taxonomic, organismal, and provenance information, creating machine-readable chemical collections whose informational value extends beyond compounds for which a specific pharmacological endpoint has already been measured [1, 2]. These resources should not be interpreted as complete or unbiased maps of natural-product

chemistry, because database membership reflects discovery practices, publication patterns, organism sampling, dereplication decisions, and curation rules. Nevertheless, they change the computational starting point: the absence of an activity label does not imply the absence of learnable chemical structure.

The asymmetry becomes clearer when chemical identity is compared with experimental activity annotation. NPASS compiles quantitative experimental activities for a large subset of natural products while also containing many additional structures without corresponding

Corresponding author: James O'Leary

Address: Department of Self-Supervised Learning for Natural Products, School of Pharmacy, Trinity College Dublin, Dublin, Ireland.

E-mail: james.oleary@tcd.ie

Received: 21 May 2025; **Revised:** 02 September 2025; **Accepted:** 09 September 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.



experimentally measured activity values in the database [3]. That observation supports a bounded claim rather than a census of all natural-product data: within an important curated resource, chemical records substantially exceed uniformly available experimental bioactivity labels. For machine learning, this means that a model trained only after endpoint annotation discards potentially informative structure until an expensive and assay-dependent label becomes available.

Experimental analytical data provide a second form of underused information. NP-MRD demonstrates that natural-product NMR spectra and associated structural information can be organized as reusable computational resources rather than remaining isolated attachments to individual structure-elucidation studies [4]. Tandem mass-spectral collections likewise contain rich chemical regularities, although their reuse depends on metadata integrity, harmonized representations, peak processing, and reproducible cleaning; the matchms library-cleaning work shows that these decisions materially affect the analytical substrate presented to downstream algorithms [5]. Spectra are therefore neither noise-free labels nor interchangeable surrogates for molecular structures, but experimentally generated views conditioned by instruments, acquisition settings, ion chemistry, and sample context.

The computational premise developed here is that natural-product learning should begin before bioactivity screening whenever chemically informative unlabelled views are available. “Before screening” does not mean replacing experiments with representation learning, nor does it imply that pretraining can recover pharmacological mechanisms absent from the data. It instead defines an ordering principle: first learn recurring chemical regularities from structures and spectra, then expose that representation to scarce task-specific biological supervision while testing whether the resulting features retain distinctions relevant to the intended discovery problem.

Why unlabelled chemical collections should be treated as training data

Self-supervision makes unlabelled molecules useful by constructing learning problems from the data themselves rather than requiring an externally assigned bioactivity endpoint. Molecular contrastive learning has shown that graph-based augmentations can produce transferable representations across molecular-property tasks, demonstrating that useful supervisory structure can be created from unlabelled molecular graphs [6]. The important premise for natural products is not that those benchmark gains automatically transfer to their broader stereochemical and scaffold diversity; it is that molecular connectivity contains regularities that can be learned before the ultimate prediction task is specified.

Molecular strings provide a complementary example. Large-scale chemical-language modelling has shown that representations trained on very large collections of unlabelled molecular sequences can capture structural and property-related information transferable to downstream tasks [7]. Scale alone, however, is not evidence of chemical understanding. Token frequencies, canonicalization conventions, scaffold recurrence, and database composition may all provide shortcuts, so the value of an unlabelled corpus depends on what the pretraining objective forces the model to reconstruct, discriminate, or align.

For natural products, the strongest argument for pretraining is therefore informational rather than numerical. A molecular graph contains connectivity; a molecular string expresses a chosen encoding of that connectivity; provenance relates structures to biological sources; and spectra record measurement-conditioned consequences of molecular composition and structure. These views differ in what they reveal and in how they can fail. The proposed framework treats them as potential sources of self-supervisory signal only when the training objective preserves chemically consequential distinctions instead of rewarding invariance to them.

The distinction among these forms of unlabelled information is summarized in **Table 1** because their roles are easy to blur when all are described simply as “pretraining data.”

Table 1. Unlabelled chemical information as distinct sources of representation-learning signal

Construct	Information available before bioactivity	Legitimate learning role	Principal boundary
Molecular structure	Connectivity and substructure relationships	Learn structural regularities	Database presence does not establish activity
Biological provenance	Structure–organism relationships	Supply contextual chemical information	Recorded provenance is sampling-dependent
Unmeasured activity space	Structures lacking experimental activity values in a curated resource	Motivate label-efficient pretraining	Missing label is not evidence of inactivity
Experimental spectra	Measurement-conditioned structural information	Learn observable chemical correspondences	Spectra are incomplete views of structure
Curated MS/MS data	Peaks plus acquisition metadata	Support spectral representation learning	Preprocessing can alter learned regularities

Molecular graphs	Topological transformations	Construct label-free contrastive tasks	Augmentations may erase relevant chemistry
Molecular strings	Sequential chemical encodings	Learn corpus-scale molecular patterns	Frequency and syntax can create shortcuts

Note: The distinction among “learning role” categories is proposed here as an organizational framework; the evidence basis for each underlying data source or training strategy is indicated by the cited study.

What should a molecular model learn before it sees bioactivity?

The central methodological question is not whether a model can be pretrained, but which invariances and distinctions should be learned before biological labels arrive. Knowledge-guided molecular pretraining shows that self-supervision can be organized around chemically informative priors rather than relying exclusively on arbitrary corruption or augmentation [8]. For natural products, this suggests that an acceptable pretext objective should force recovery or comparison of information whose loss would matter chemically, while avoiding tasks solvable through database identifiers, trivial atom-count correlations, representation syntax, or other features disconnected from the intended molecular reasoning.

Functional groups provide one possible level of supervision. FG-BERT demonstrates that functional-group information can structure self-supervised molecular learning and improve downstream property prediction in the studied benchmarks [9]. Yet functional groups should not be treated as complete pharmacophoric explanations: identical motifs can behave differently because of stereochemistry, substitution pattern, conformational restriction, electronic context, macrocyclization, glycosylation, or their placement within a larger scaffold. The framework therefore treats functional groups as chemically interpretable intermediate units rather than privileged final units of meaning.

External chemical relations provide another source of inductive structure. Knowledge-graph-enhanced contrastive learning has demonstrated that explicit chemical knowledge can be coupled to molecular representations rather than leaving all similarity relations to emerge from topology alone [10]. Such knowledge remains selective and potentially biased; a knowledge graph encodes what has been recorded, not an unbiased chemical ontology. The proposed criterion is therefore chemical consequence preservation: pretraining should preferentially retain molecular differences that plausibly change structure, reactivity, recognition, or experimental observables, while treating any imposed invariance as a hypothesis that must later be challenged.

This criterion also prevents a common conceptual mistake. A representation is not chemically meaningful merely because it compresses molecules efficiently or improves one supervised endpoint. Before bioactivity fine-tuning, its desirable content is more modest: sensitivity to chemically

consequential perturbations, organization across multiple structural scales, compatibility with experimentally observed views, and enough flexibility to support later task-specific distinctions without assuming them in advance.

Self-supervision from molecular graphs

Molecular graphs provide an attractive substrate because atoms and bonds supply an explicit relational representation while allowing labels to be generated from transformations of the molecule itself. Contrastive graph learning demonstrates that differently augmented views of a molecular graph can be brought together in representation space and subsequently transferred to property-prediction tasks [6]. For natural products, however, the critical design decision is whether an augmentation creates two chemically equivalent views or destroys precisely the motif that later governs conformation, reactivity, recognition, or fragmentation. Random graph perturbation is therefore a computational operation, not automatically a chemically valid invariance. Functional-group-aware pretraining illustrates a more constrained strategy because it asks the learner to exploit recognizable chemical organization rather than treating every subgraph as semantically interchangeable [9]. A natural-product implementation could extend this principle to ring systems, glycosidic attachments, oxidation patterns, heteroatom environments, stereochemical centres, or scaffold–substituent relationships, but those extensions are proposed rather than established by the cited work. Their suitability would have to be tested against tasks in which preserving the nominated distinction can be independently evaluated.

Knowledge-guided graph objectives further show that pretraining can integrate information beyond local connectivity. KPGT couples molecular graph learning with chemically relevant knowledge during pretraining [8], whereas KANO demonstrates that chemical-knowledge relationships can be incorporated into contrastive molecular learning [10]. Together, these studies justify rejecting a false choice between purely unsupervised topology and fully endpoint-supervised prediction. The proposed natural-product framework instead places graph pretraining on a continuum: topology supplies the base representation, chemically motivated pretext tasks constrain what transformations are acceptable, and

external knowledge can regularize the latent space without being treated as infallible ground truth.

Graph self-supervision therefore serves two purposes before screening. It can expose the model to structural diversity beyond the labelled assay set, and it can test whether chemically related transformations are represented consistently. It cannot, by itself, establish that two similar graphs share pharmacology, that a latent neighbourhood represents mechanism, or that the learned invariances remain valid under stereochemical, conformational, or experimental domain shifts.

Self-supervision from three-dimensional structure

Two-dimensional connectivity does not uniquely specify the spatial state presented to a protein, solvent, membrane, chromatographic system, or spectrometer. Geometry-enhanced molecular representation learning demonstrates that explicit geometric information can complement graph topology in downstream molecular prediction [11]. The implication for natural-product pretraining is especially important for conformationally constrained rings, stereochemically dense scaffolds, polycyclic systems, and molecules whose intramolecular interactions reorganize accessible shape. Geometry should therefore be represented as additional evidence about molecular state rather than as decorative coordinates appended to a graph. At the same time, three-dimensional data introduce new uncertainty. A deposited conformer may reflect a computational minimum, a crystal environment, a particular protonation state, or one member of a broader ensemble; treating it as the molecule's single "true" geometry would create false precision. The proposed framework consequently distinguishes connectivity invariance from state-conditioned geometry: transformations that preserve graph identity need not preserve pharmacologically or spectroscopically relevant conformation, and pretraining should not force geometrically distinct states together without a chemical justification.

Recent multi-channel molecular learning also supports the broader principle that structural information can be organized across multiple levels rather than flattened into a single atom-wise representation [12]. Structural hierarchy is useful here because natural-product meaning may arise from interactions among local motifs, ring systems, substituent patterns, and global molecular shape. The framework therefore proposes hierarchical geometric pretraining in which local and global relationships can coexist, while refusing to infer that any particular hierarchy is universally optimal.

A chemically meaningful 3D representation should ultimately survive tests beyond conventional property prediction: stereoisomer discrimination where

stereochemistry changes observables, sensitivity to conformationally relevant perturbations, robustness to alternative plausible conformers, and compatibility with experimental modalities. Those tests remain validation requirements of the proposed framework, not outcomes already demonstrated for natural-product discovery.

Spectral self-supervision as experimental chemical knowledge

Spectra offer a fundamentally different pretraining substrate because they are observations produced when molecules interact with a measurement process. Learned MS/MS representations have already shown that spectral comparison can move beyond literal peak overlap toward data-driven similarities associated with molecular structural relationships [13, 14]. These methods should not be collapsed into one learning category: Spec2Vec derives unsupervised spectral embeddings, whereas MS2DeepScore learns a neural similarity function using structural relationships. Their joint relevance is narrower but important—tandem spectra contain regularities from which chemically useful representations can be learned.

Contrastive learning provides a more direct bridge to self-supervision. CLERMS demonstrates that contrastive objectives can be used to construct tandem-mass-spectral embeddings [15]. For natural-product collections, such objectives could exploit replicate spectra, acquisition variants, structurally related spectra, or paired molecular information, but positive and negative construction becomes a scientific decision: two spectra may differ because of collision energy, adduct chemistry, instrument platform, ionization conditions, isomerism, or genuine structural change. A model that ignores these distinctions may learn instrument invariance successfully while simultaneously erasing chemical information.

Domain-informed spectral prediction supplies complementary evidence that measurement data are chemically structured rather than arbitrary high-dimensional signals. Chemical-formula transformers have used chemistry-inspired constraints to annotate metabolite mass spectra [16], while graph-transformer models can learn structure-conditioned tandem-spectrum prediction [17]. These achievements do not mean that formula prediction resolves structure, that forward spectral prediction uniquely solves the inverse problem, or that learned fragmentation patterns constitute a proven mechanistic model. They show instead that composition, structure, and measured spectra share learnable statistical relationships that can discipline representation learning.

Natural-product structure annotation makes the case for preserving modality specificity. Machine-learning-assisted analytical workflows already operate across MS and NMR information in natural-product structure elucidation, but

their evidence base remains heterogeneous in modality, chemical class, database dependence, and validation setting [18]. Spectral pretraining should therefore retain what is experimentally distinctive about each modality while preparing representations for later cross-modal alignment. The objective is not to convert every spectrum into an allegedly complete molecular identity, but to learn which experimental regularities can inform molecular representation before scarce bioactivity labels are introduced.

Connecting molecular and spectral views of the same compound

A structure and its spectrum describe the same compound from different observational positions. Contrastive mass-spectrum–structure pretraining demonstrates that paired tandem spectra and molecular structures can be embedded into a shared space for metabolite identification [19]. Cross-modal retrieval between ^{13}C NMR spectra and structures similarly shows that experimentally derived spectral information can be aligned with molecular representations for candidate retrieval [20]. These results support alignment as a computational operation, but not the stronger claim that one modality completely determines another.

The distinction matters because shared chemical information coexists with modality-specific information. Infrared spectrum–structure contrastive learning extends the feasibility of cross-modal alignment to another spectroscopy, although its simulated, comparatively small-molecule setting limits direct extrapolation to structurally complex natural products [21]. Joint embedding of tandem spectra and candidate molecules can also support metabolite ranking [22], yet a correctly ranked candidate is not thereby shown to possess a particular bioactivity or mechanism. The proposed framework therefore treats cross-modal agreement as evidence of correspondence, not equivalence.

For natural products, alignment should preserve three possibilities: two views may contain mutually recoverable chemical information; one view may contain information absent from the other; or apparent disagreement may reflect acquisition conditions, conformational state, stereochemistry, annotation error, or incomplete candidate coverage. This distinction prevents the latent space from being optimized merely to erase experimentally meaningful disagreement. **Table 2** organizes these cross-modal relationships, and **Figure 1** expresses the resulting architecture visually.

Table 2. Conceptual distinctions for aligning molecular structure with experimental spectra

Relationship	Interpretation	Appropriate representation goal	Boundary
Shared signal	Both views encode related chemical structure [19]	Align recoverable correspondence	Alignment is not identity
Structure-specific signal	Molecular representation contains information not resolved by a given spectrum [20]	Retain rather than force agreement	Retrieval remains candidate-dependent
Spectrum-specific signal	Measurement contains condition-dependent observable patterns [21]	Preserve modality context	Simulated or instrument-specific spectra may shift
Task-facing correspondence	Joint embeddings support candidate ranking [22]	Transfer aligned features downstream	Ranking does not establish bioactivity
Discordance	Views disagree beyond expected correspondence	Flag rather than collapse	May reflect chemistry, measurement, or annotation

Note: The five-way distinction and representation goals are proposed here; cited studies establish the underlying feasibility and boundaries of cross-modal learning.

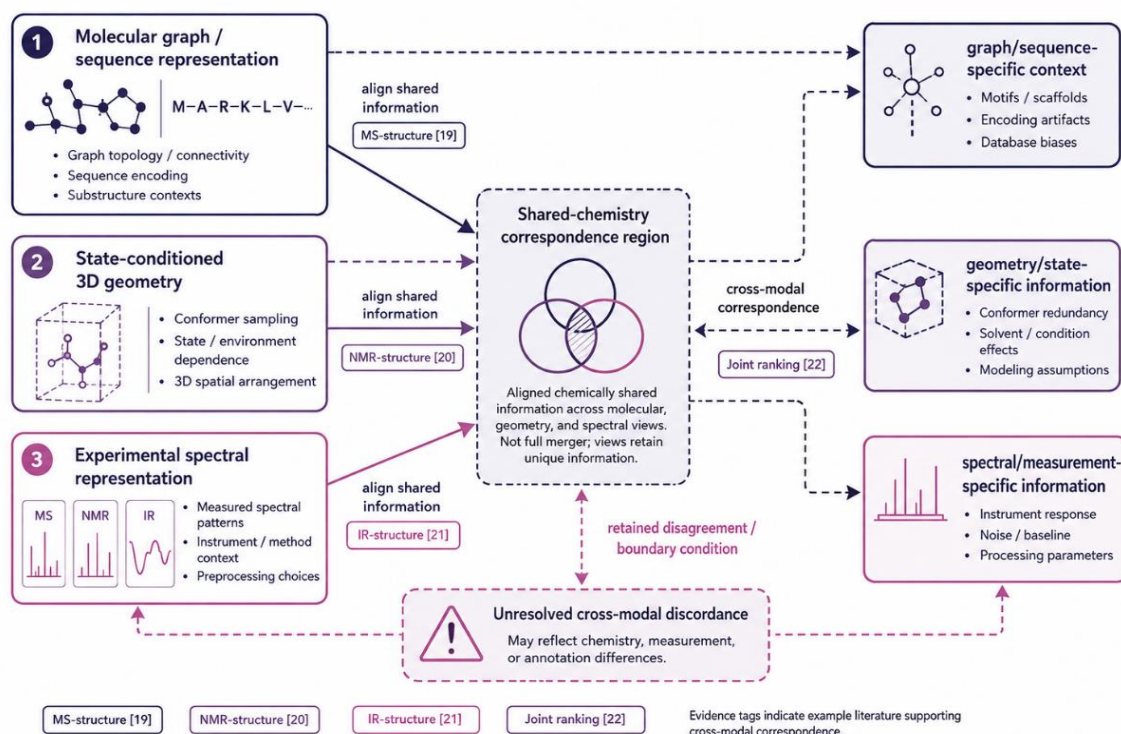


Figure 1. Correspondence without collapse in natural-product multiview pretraining

A multiview natural product pretraining framework

The proposed framework combines these premises into a staged representation strategy rather than a single universal encoder. Complementary molecular modalities can benefit from non-overlapping self-supervised learning, supporting the idea that different representations need not be treated as redundant copies of the same information [23]. In the natural-product setting proposed here, graph or sequence views establish connectivity-level context, geometry contributes state-dependent spatial information, and spectra introduce experimentally observed chemical consequences. Shared latent components are encouraged where evidence supports correspondence, whereas modality-specific components are retained when forcing agreement would remove information.

Pretraining would therefore alternate among within-view objectives and cross-view objectives. Within-view tasks should reconstruct or discriminate chemically consequential features; cross-view tasks should align paired observations only to the degree justified by their shared chemistry. Missing spectra, uncertain conformers, or incomplete provenance should not require synthetic completeness. Instead, the architecture is proposed to permit partially observed compounds to contribute through the views actually available.

Representation agreement should also be asymmetric when the data warrant it. A clean structural view may help interpret a noisy spectrum without requiring the reverse mapping to be equally informative, while spectra acquired

under different conditions may reveal which molecular features remain observable. Agreement is therefore better treated as a graded relation than a binary constraint.

The framework is deliberately upstream of biological interpretation. It does not assume that an embedding coordinate corresponds to potency, target engagement, mechanism, exposure, or therapeutic effect. Its purpose is narrower: produce a representation that can later receive scarce labels without having learned chemistry only through the narrow distribution of a particular screen.

Avoiding pretext tasks that reward trivial chemical shortcuts

A pretext task can be solved well for the wrong reason. Chemical-synthesis knowledge has been used to constrain contrastive molecular representation learning, showing that chemically structured relationships can be introduced rather than leaving positive-pair construction arbitrary [24]. For natural products, however, synthetic reaction knowledge cannot simply be interpreted as biosynthetic truth. The broader principle is that positive pairs, masking rules, and invariances should be defensible in chemical terms.

The proposed shortcut audit asks whether success remains possible after trivial correlates are neutralized. Models should be challenged against scaffold frequency, molecular size, representation syntax, acquisition platform, library identity, and other features that may predict the pretext label without requiring the intended chemistry. A

useful objective should also fail appropriately when a perturbation changes stereochemistry, functional context, geometry, or spectral evidence that the downstream problem requires.

The same audit should distinguish nuisance invariance from destructive invariance. Removing a file identifier or instrument-specific formatting artifact may improve robustness, whereas making stereoisomers, regioisomers, or distinct adduct states indistinguishable can eliminate information needed later. The appropriate invariance is therefore task-contingent but chemically bounded: it should be justified by the information that the downstream assay or analytical process can reasonably depend upon.

Shortcut resistance is therefore not another accuracy metric; it is a test of what information the representation was forced to use.

Fine-tuning for rare bioactivity labels

Pretraining becomes discovery-relevant only when representations can be adapted to a defined biological question. Natural-product LC-MS/MS features have supported supervised bioactivity classification in a bounded experimental setting [25], establishing that analytical molecular information can be connected computationally to activity labels without implying target mechanism or therapeutic efficacy. The proposed framework uses such labels only after chemistry-focused pretraining.

Fine-tuning should preserve endpoint identity. Binding, functional response, phenotypic activity, cytotoxicity, and organism-level efficacy are not interchangeable labels, and a representation useful for one need not transfer to another. With rare labels, the preferred test is not whether full fine-tuning can memorize the available assay set, but whether pretrained features improve data efficiency under chemically demanding splits while remaining calibrated. Any claimed benefit would require comparison against non-pretrained and simpler representation baselines using identical training and evaluation conditions.

Evaluating whether learned representations are chemically meaningful

Molecular benchmarks formalize datasets, tasks, splits, and metrics, making explicit that representation quality cannot be judged independently of evaluation design [26]. A systematic study of molecular-property prediction further shows that observed performance depends on interacting choices in representation, architecture, training, and data [27]. Consequently, downstream accuracy alone cannot establish that pretraining has captured chemically meaningful structure.

The framework proposes a broader evaluation battery. Structural probes should test whether local neighborhoods

remain coherent under chemically conservative changes yet separate when stereochemistry, scaffold organization, or functional context changes materially. Cross-modal probes should assess whether paired spectra and structures retrieve one another while preserving modality-specific information. Label-efficiency tests should compare performance as biological supervision becomes sparse, and ablations should isolate whether gains arise from graph, geometry, spectrum, or alignment objectives.

Chemical-meaning evaluation should additionally separate representation diagnostics from downstream utility. A model may preserve stereochemical distinctions yet offer no advantage for an assay insensitive to stereochemistry, or improve an assay endpoint through correlations that are chemically uninformative. Reporting both layers prevents a favorable downstream result from retroactively defining every feature in the embedding as meaningful chemistry.

These probes remain diagnostics rather than proofs of mechanism. A latent representation may correlate with structural or activity distinctions because of dataset composition, and chemically intuitive neighborhoods can still encode hidden sampling biases. Meaningfulness should therefore be treated as convergent evidence from several controlled tests, not as a property inferred from one benchmark score.

Transfer across scaffold, target, and experimental domains

Generalization is especially vulnerable when training and evaluation occupy overlapping chemical space. Coverage bias can materially influence small-molecule machine-learning performance, making explicit applicability analysis necessary when claiming transfer beyond familiar chemistry [28]. For natural products, scaffold-held-out evaluation should therefore be complemented by tests across provenance, spectral acquisition conditions, and underrepresented structural classes.

Transfer is also relational rather than automatically beneficial. Transferability mapping shows that usefulness depends on the source–target pairing rather than on pretraining being universally advantageous [29]. The proposed framework consequently treats transfer across scaffolds, targets, assays, and experimental platforms as separate questions. Negative transfer is an informative failure mode: it may indicate that pretraining encoded the wrong invariances, that source and target domains differ materially, or that the downstream labels require information absent from the pretrained views.

Data boundaries and representation failure

Natural-product pretraining inherits the boundaries of its source collections. Provenance databases reflect what organisms and compounds have been recorded [2], while mass-spectral libraries remain sensitive to cleaning,

metadata, and acquisition differences [5]. Analytical machine-learning evidence in natural-product structure annotation is likewise heterogeneous across modalities and validation settings [18]. A large corpus can therefore amplify systematic coverage patterns as readily as it can broaden chemical exposure.

Representation failure should be diagnosed rather than hidden inside aggregate metrics. Chemical-space coverage may produce apparently strong interpolation while leaving remote scaffolds poorly represented [28]. Source–target mismatch may also make transfer neutral or harmful [29]. Additional failures can arise when conformers are treated as fixed truth, spectra are stripped of relevant acquisition context, cross-modal objectives force incompatible observations together, or rare biological labels redefine the task in ways pretraining could not anticipate.

Failure analysis also needs a missing-data perspective. Natural-product records rarely provide all modalities uniformly, so compounds with structures but no spectra, spectra but uncertain structures, or incomplete provenance should not be silently discarded. The proposed architecture treats missingness itself as a boundary condition and requires evaluation to determine whether performance differences reflect chemistry, modality availability, or systematic patterns in which compounds receive richer characterization.

The framework therefore has an explicit abstention boundary: when a compound, spectrum, scaffold, or assay lies outside demonstrated representation coverage, confidence should not be inferred from embedding proximity alone. Prospective natural-product validation remains necessary before the proposed architecture can be regarded as discovery-effective.

CONCLUSION

Natural-product machine learning need not begin at the point where a scarce biological label becomes available. Structures, molecular encodings, geometry, provenance, and experimental spectra provide complementary opportunities to learn chemical regularities before task-specific screening. The framework proposed here organizes those opportunities around chemical consequence preservation, modality-aware self-supervision, correspondence without forced equivalence, shortcut testing, label-efficient adaptation, and explicit transfer boundaries.

Its central claim is methodological rather than predictive: representation learning should be judged by what chemistry it preserves and what uncertainty it exposes, not only by whether pretraining increases one downstream score. A useful natural-product representation should remain sensitive to consequential structural changes,

tolerate justified nuisance variation, connect experimental and molecular views where correspondence exists, and retain disagreement when it does not.

This framework is not a validated universal architecture and does not establish biological mechanism, therapeutic efficacy, or prospective discovery performance. Its value depends on rigorous comparisons, chemically challenging splits, multimodal ablations, calibration, domain-shift assessment, and prospective evaluation on genuinely unseen natural products. Learning chemistry before learning the screen is therefore best treated as a testable computational strategy for using abundant chemical information more intelligently.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] 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/gkac1063
- [2] 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
- [3] Zhao H, Yang Y, Wang S, Yang X, Zhou K, Xu C, et al. NPASS database update 2023: Quantitative natural product activity and species source database for biomedical research. *Nucleic Acids Res.* 2023;51(D1). doi:10.1093/nar/gkac1069
- [4] Wishart DS, Sayeeda Z, Budinski Z, Guo AC, Lee BL, Berjanskii M, et al. NP-MRD: The Natural Products Magnetic Resonance Database. *Nucleic Acids Res.* 2022;50(D1). doi:10.1093/nar/gkab1052
- [5] de Jonge NF, Hecht H, Strobel M, Wang M, van der Hooft JJJ, Huber F. Reproducible MS/MS library cleaning pipeline in matchms. *J Cheminform.* 2024;16(1):88. doi:10.1186/s13321-024-00878-1
- [6] 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
- [7] Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and

- properties. *Nat Mach Intell.* 2022;4(12):1256-64. doi:10.1038/s42256-022-00580-7
- [8] 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
- [9] Li B, Lin M, Chen T, Wang L. FG-BERT: A generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Brief Bioinform.* 2023;24(6). doi:10.1093/bib/bbad398
- [10] Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nat Mach Intell.* 2023;5(5):542-53. doi:10.1038/s42256-023-00654-0
- [11] 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
- [12] 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
- [13] Huber F, Ridder L, Verhoeven S, Spaaks JH, Diblen F, Rogers S, et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput Biol.* 2021;17(2). doi:10.1371/journal.pcbi.1008724
- [14] Huber F, van der Burg S, van der Hooft JJJ, Ridder L. MS2DeepScore: A novel deep learning similarity measure to compare tandem mass spectra. *J Cheminform.* 2021;13(1):84. doi:10.1186/s13321-021-00558-4
- [15] Guo H, Xue K, Sun H, Jiang W, Pu S. Contrastive learning-based embedder for the representation of tandem mass spectra. *Anal Chem.* 2023;95(20):7888-96. doi:10.1021/acs.analchem.3c00260
- [16] Goldman S, Wohlwend J, Stražar M, Haroush G, Xavier RJ, Coley CW. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat Mach Intell.* 2023;5(9):965-79. doi:10.1038/s42256-023-00708-3
- [17] Young A, Röst H, Wang B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nat Mach Intell.* 2024;6(4):404-16. doi:10.1038/s42256-024-00816-8
- [18] Hu G, Qiu MH. Machine learning-assisted structure annotation of natural products based on MS and NMR data. *Nat Prod Rep.* 2023;40(11):1735-53. doi:10.1039/D3NP00025G
- [19] Chen L, Xia B, Wang Y, Huang X, Gu Y, Wu W, et al. CMSSP: A contrastive mass spectra-structure pretraining model for metabolite identification. *Anal Chem.* 2024;96(42):16871-81. doi:10.1021/acs.analchem.4c03724
- [20] Yang Z, Song J, Yang M, Yao L, Zhang J, Shi H, et al. Cross-modal retrieval between ^{13}C NMR spectra and structures for compound identification using deep contrastive learning. *Anal Chem.* 2021;93(50):16947-55. doi:10.1021/acs.analchem.1c04307
- [21] Kanakala GC, Sridharan B, Priyakumar UD. Spectra to structure: Contrastive learning framework for library ranking and generating molecular structures for infrared spectra. *Digit Discov.* 2024;3(12):2417-23. doi:10.1039/D4DD00135D
- [22] Kalia A, Chen YZ, Krishnan D, Hassoun S. JESTR: Joint embedding space technique for ranking candidate molecules for the annotation of untargeted metabolomics data. *Bioinformatics.* 2025;41(7). doi:10.1093/bioinformatics/btaf354
- [23] 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
- [24] Xie J, Wang Y, Rao J, Zheng S, Yang Y. Self-supervised contrastive molecular representation learning with a chemical synthesis knowledge graph. *J Chem Inf Model.* 2024;64(6):1945-54. doi:10.1021/acs.jcim.4c00157
- [25] Brittin NJ, Anderson JM, Braun DR, Rajski SR, Currie CR, Bugni TS. Machine learning-based bioactivity classification of natural products using LC-MS/MS metabolomics. *J Nat Prod.* 2025;88(2):361-72. doi:10.1021/acs.jnatprod.4c01123
- [26] 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
- [27] Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. *Nat Commun.* 2023;14(1):6395. doi:10.1038/s41467-023-41948-6
- [28] Kretschmer F, Seipp J, Ludwig M, Klau GW, Böcker S. Coverage bias in small molecule machine learning. *Nat Commun.* 2025;16(1):554. doi:10.1038/s41467-024-55462-w
- [29] Yao S, Song J, Jia L, Cheng L, Zhong Z, Song M, et al. Fast and effective molecular property prediction with transferability map. *Commun Chem.* 2024;7(1):85. doi:10.1038/s42004-024-01169-4