Molecular Foundation Models for Phytochemical Space: Encoding Natural Product Structure, Bioactivity, Target Context, and Therapeutic Potential

Mateo Alvarez[1]*, Sofia Herrera[2], Diego Cruz[1]

[1]Department of AI for Amazonian Shamanic Plant Extracts, Faculty of Pharmacy, National University of Colombia, Bogota, Colombia.

[2]Department of Computational Biophysics of Peptide Toxins, Faculty of Pharmacy, University of Chile, Santiago, Chile.

ABSTRACT

Molecular foundation models are emerging as general-purpose representation systems for chemical and biological prediction, yet phytochemical space remains poorly served by models trained mainly on generic small-molecule data. This article proposes a conceptual framework for molecular foundation models designed specifically to represent natural product complexity without converting computational predictions into unsupported therapeutic claims. The framework organizes phytochemical information across identity, source, chemical structure, stereochemistry, scaffold architecture, biosynthetic context, molecular graphs, molecular language, three-dimensional information where supported, bioactivity evidence, assay context, target and pathway relationships, safety signals, provenance, data quality, and uncertainty. It distinguishes representation learning from evidence interpretation by separating pretraining, transfer learning, multimodal integration, and fine-tuning from the validation steps required to establish bioactivity, target engagement, safety, or therapeutic usefulness. Target-context encoding is positioned as a means of generating testable hypotheses and prioritizing experiments, while therapeutic potential is treated as a bounded research construct rather than a model-confirmed property. The proposed design requirements emphasize natural product-aware training data, stereochemistry-sensitive encoders, assay-aware labels, provenance tracking, calibrated uncertainty, bias assessment, external validation, experimental confirmation, expert curation, and transparent reporting. The main contribution is an original, evidence-grounded architecture for aligning molecular foundation model development with the distinctive informational structure of phytochemical space and with responsible translational interpretation.

Key Words: *Molecular foundation models, Phytochemical space, Natural products, Molecular representation, Bioactivity prediction, Target context*

eIJPPR 2026; 16(1):71-87

HOW TO CITE THIS ARTICLE: Alvarez M, Herrera S, Cruz D. Molecular Foundation Models for Phytochemical Space: Encoding Natural Product Structure, Bioactivity, Target Context, and Therapeutic Potential. Int J Pharm Phytopharmacol Res. 2026;16(1):71-87. https://doi.org/10.51847/nk0XqwaEWk

INTRODUCTION

Molecular foundation models extend conventional chemical representation learning by using broad pretraining objectives to construct reusable representations that can subsequently support task-specific adaptation. Their potential value in drug discovery lies not in an assumed ability to understand chemistry, but in their capacity to organize recurring structural patterns and transfer learned representations across related prediction tasks. This distinction is especially important because a molecule can be encoded through descriptors, fingerprints, molecular strings, graphs, geometric features, or combinations of these modalities, each preserving different

Corresponding author: Mateo Alvarez
Address: Department of AI for Amazonian Shamanic Plant Extracts, Faculty of Pharmacy, National University of Colombia, Bogota, Colombia.
E-mail: ✉ mateo.alvarez@unal.edu.co
Received: 10 November 2025; **Revised:** 14 February 2026; **Accepted:** 16 February 2026

information and imposing different inductive assumptions. Consequently, the usefulness of a molecular foundation model depends on whether its representation is chemically appropriate for the intended problem and whether its outputs remain interpretable within the limits of the encoded data [1].

Evaluation presents a second foundational challenge. Molecular models are commonly assessed across property, bioactivity, toxicity, and interaction tasks, but apparent performance depends strongly on the selected dataset, endpoint definition, data split, metric, and chemical similarity between development and evaluation compounds. A model that performs well under a random split may not transfer to new scaffolds, uncommon stereochemical configurations, underrepresented compound classes, or natural products collected from poorly represented taxa. Standardized molecular-learning benchmarks therefore provide useful comparative infrastructure, but benchmark success alone cannot demonstrate that a representation generalizes to phytochemical space or supports pharmacologically meaningful interpretation [2].

These limitations are amplified for natural products. Phytochemicals occupy structurally and informationally heterogeneous regions of chemical space characterized by complex ring systems, dense stereochemistry, glycosylation, macrocyclic architectures, unusual functional-group combinations, and biosynthetically related analogue families. Their activity records may also be sparse, selectively reported, derived from incompatible assay formats, or concentrated around frequently investigated compounds and targets. Comparative analyses of molecular-property prediction show that performance is shaped not only by model architecture but also by dataset size, molecular representation, activity cliffs, label quality, and evaluation strategy [3]. A phytochemical foundation model must therefore encode data quality and applicability limits rather than treating all structures and activity labels as equally informative.

The pharmacological interpretation of model outputs introduces a further boundary. Predicted bioactivity does not establish experimentally observed activity; predicted compound–target association does not confirm binding or functional engagement; pathway proximity does not demonstrate causal mechanism; and favourable computational safety outputs do not establish tolerability. Machine learning may assist compound organization, hypothesis generation, target prioritization, and experimental planning across defined stages of drug discovery, but these functions remain dependent on suitable evidence, domain expertise, and empirical validation [4]. The need is therefore not merely for a larger molecular model, but for a phytochemical-space framework that explicitly integrates structure, stereochemistry, source context, bioactivity evidence, target and pathway context, safety information, provenance, uncertainty, and translational boundaries.

Phytochemical space and molecular foundation models

Phytochemical space can be defined as the multidimensional domain formed by natural product identities, chemical structures, stereochemical configurations, scaffold families, compound classes, biological sources, biosynthetic relationships, measured activities, target associations, safety observations, and evidence provenance. It is broader than a structure-only collection because the same nominal compound may be represented differently across databases, linked to multiple organisms, reported with incomplete stereochemistry, or associated with conflicting biological observations. Reviews of natural product databases show substantial differences in coverage, annotation practices, structure standardization, source information, and linkage to biological data [5]. A natural product-aware foundation model must therefore begin with record reconciliation and context preservation rather than assuming that aggregated database entries are interchangeable training examples.

Within this domain, a molecular foundation model is proposed here as a broadly pretrained representation system that can be adapted to several phytochemical research tasks without claiming universal validity across them. Its pretraining foundation may include molecular graphs, molecular strings, descriptors, chemical classes, source annotations, spectra, literature-derived relationships, target information, or other modalities, provided that each input retains an identifiable evidentiary status. Large open collections of natural products can provide structural diversity for such pretraining, but their primary value is coverage rather than automatic pharmacological annotation. Aggregated natural product resources contain standardized structures and associated descriptors that can support chemical-space exploration, representation learning, and retrieval, while incomplete source or activity metadata still require qualification [6]. Pretraining coverage should therefore be documented across scaffolds, compound classes, stereochemical states, and source categories.

Natural product-aware representation also requires clarity about the different roles of molecular language models, graph foundation models, self-supervised learning, transfer learning, and multimodal integration. A molecular language model learns statistical relationships among tokens in a molecular notation; a graph foundation model operates on atoms, bonds, and graph-derived relationships; and self-supervised learning constructs training objectives from the molecular data themselves rather than relying

exclusively on labelled endpoints. Transfer learning adapts a pretrained representation to a downstream task, whereas multimodal learning aligns or integrates different information types. None of these approaches is inherently natural product-aware. Even transparent fixed representations such as chemical fingerprints emphasize different aspects of molecular similarity, and their effectiveness in natural product space varies with fingerprint design and the chemical relationships being examined [7]. A phytochemical model should consequently compare learned embeddings against multiple auditable baselines.

The distinction between generic small-molecule space and phytochemical space is not simply one of molecular complexity. Natural product records may carry botanical, microbial, marine, or other organismal origins; taxonomic identifiers; tissue or material information; biosynthetic clues; compound-class assignments; literature provenance; and varying levels of curation. These fields can influence whether apparently identical records should be merged, kept separate, or assigned different confidence levels. Curated natural product resources increasingly preserve structural standardization, stereochemical information, molecular classifications, source links, and processing histories, while acknowledging that residual annotation errors and uncertain natural-product status may remain [8]. In the proposed framework, source and provenance metadata are therefore represented as model-accessible context and as governance information that can constrain interpretation, rather than as decorative annotations added after prediction.

Structure and bioactivity encoding

Structure encoding should combine complementary representations rather than assuming that one molecular view captures every feature relevant to phytochemical space. Molecular-language pretraining can learn transferable patterns from large corpora of molecular strings and support later adaptation to property-prediction tasks, but its outputs remain dependent on tokenization, molecular notation, corpus composition, and downstream evaluation [9]. Graph-based contrastive learning offers a different route by constructing representations through atom–bond topology and self-supervised molecular augmentations [10]. For phytochemicals, such augmentations require chemical scrutiny because deletion of a bond, subgraph, stereochemical marker, glycosidic linkage, or rare substituent could remove precisely the feature that distinguishes a natural product scaffold or controls its biological recognition.

Stereochemistry and geometry require dedicated treatment. A structure-processing pipeline should preserve defined stereocentres, double-bond geometry, undefined stereochemistry, charges, tautomeric decisions, and structure-normalization history. Geometry-enhanced representation learning demonstrates that bond lengths, bond angles, and related spatial features can improve molecular-property representations when suitable geometric information is available [11]. Nevertheless, generated conformers should be distinguished from experimentally observed conformations, and neither should automatically be equated with a biologically active conformation. Two-dimensional molecular depictions can also support self-supervised representation and downstream property or target-prediction tasks, illustrating that image-like molecular views may provide complementary signals [12]. In the proposed framework, molecular images and three-dimensional information are optional evidence-qualified modalities rather than replacements for stereochemistry-aware graphs and molecular strings.

Cross-modal alignment can reduce dependence on the limitations of a single representation. Contrastive learning between molecular graphs and molecular strings provides a mechanism for learning shared features while retaining distinct encoders for different molecular views [13]. Multimodal chemical pretraining can extend this principle by aligning structural and language-related information within a common representation space [14]. For phytochemical applications, however, a common embedding should not erase the origin of each input. The representation should preserve modality-level provenance so that a structural feature, taxonomic annotation, assay result, literature-derived relationship, or predicted target association can be inspected independently. Compound classes, scaffold features, chemical fingerprints, and natural product source should similarly be encoded as traceable contextual attributes rather than collapsed into an opaque vector with no accessible evidentiary lineage.

Bioactivity encoding requires stricter separation between observation and inference. Quantitative measurements should retain endpoint identity, assay format, species, target construct, units, relation operators, concentration conditions, controls, replicate information, and publication source wherever available. Predicted activities should be stored as model outputs rather than merged with observed labels, while uncertain, contradictory, or incompletely reported records should remain explicitly marked. Graph foundation models that combine structural pretraining with biological multitask information illustrate how pretraining and downstream adaptation can be connected across molecular-property tasks [15]. In a phytochemical framework, this connection should be accompanied by assay-aware fine-tuning, external evaluation, uncertainty reporting, and clear identification of the evidence supporting each output. **Table 1** summarises structure and

bioactivity encoding requirements for molecular foundation models representing phytochemical space.

Table 1. Structure and Bioactivity Encoding Requirements for Molecular Foundation Models in Phytochemical Space: Chemical Structure, Stereochemistry, Molecular Graphs, Molecular Language, Scaffold Features, Natural Product Source, Assay Context, Bioactivity Evidence, Provenance, and Uncertainty

Encoding dimension	Core representation question	Required input data	Foundation model function	Natural product-specific challenge	Validation requirement	Failure risk	Framework role
Chemical structure	Does the representation preserve molecular identity, connectivity, bond type, charge, aromaticity, and normalization decisions?	Standardized structure records, atom and bond annotations, charge states, tautomer decisions, and stable identifiers	Learn reusable atom-, bond-, fragment-, and molecule-level representations	Salts, mixtures, glycosides, unusual valence states, duplicate records, and inconsistent structures may be present	Structure-standardization audit, duplicate analysis, and expert review of representative records	Incorrect molecular identity propagates into every downstream task	Core identity and structural representation layer
Stereochemistry	Are defined, undefined, conflicting, and source-dependent stereochemical assignments represented explicitly?	Isomeric molecular notation, stereochemical bond labels, configuration annotations, and evidence source	Distinguish stereoisomers and encode uncertainty where configuration is unresolved	Biological recognition may depend on stereochemistry, while many database records are incomplete	Stereoisomer challenge sets and explicit evaluation of undefined-stereochemistry cases	Distinct phytochemicals collapse into misleadingly similar representations	Mandatory natural product-aware encoding control
Molecular graph	Does graph construction retain chemically meaningful topology and local attributes?	Atom, bond, ring, aromaticity, charge, and stereochemical features	Graph pretraining, message passing, substructure learning, and task adaptation	Fused rings, bridged systems, macrocycles, glycosides, and uncommon scaffolds increase representational complexity	Scaffold-aware external evaluation and chemically reviewed graph augmentations	Topological shortcuts or destructive augmentations obscure rare structural features	Principal structural encoder
Molecular language	Does molecular tokenization preserve meaningful chemical distinctions and remain robust to alternative valid strings?	Standardized isomeric strings, tokenization rules, and alternative molecular-string enumerations	Masked-token, denoising, or related self-supervised pretraining	Long strings, rare tokens, non-unique notation, and underrepresented natural product syntax	String-randomization robustness and consistency checks against graph representations	The model learns notation frequency rather than transferable chemistry	Complementary sequence-based encoder

ISSN (Online) 2249-6084 (Print) 2250-1029

Scaffold features and chemical fingerprints	Which fixed representations provide interpretable baselines and neighbourhood definitions?	Scaffold assignments, ring systems, fragments, functional groups, fingerprint families, and similarity metrics	Baseline modeling, chemical-space mapping, retrieval, and coverage assessment	Different fingerprints emphasize different aspects of natural product similarity	Multi-fingerprint comparison, scaffold splitting, and novel-scaffold testing	A single representation gives a distorted account of phytochemical similarity	Transparent baseline and generalization-audit layer
Compound class	Can hierarchical or overlapping chemical classes be represented without being treated as proof of biological function?	Curated class labels, ontology relationships, and assignment provenance	Auxiliary learning, class-conditioned retrieval, and coverage analysis	Class assignments may be inconsistent, overlapping, or derived from structure-prediction rules	Agreement testing against curated assignments and expert review	Circular class labels become proxies for activity or target labels	Semantic chemical-context layer
Natural product source and biosynthetic context	Is the compound linked to producing organism and biosynthetic information only at the strength supported by evidence?	Taxonomic identifiers, organism part or material, collection context, biosynthetic genes or pathways where available, and literature source	Align chemical and biological-source modalities and support context-aware retrieval	Taxonomic revisions, synonyms, contamination, ecological variation, and incomplete biosynthetic knowledge	Taxonomic normalization, source–compound lineage checks, and organism-held-out evaluation	False source or pathway associations distort biological hypotheses	Source and optional biosynthetic-context layer
Three-dimensional information	Is spatial information represented as an evidence-qualified conformational hypothesis?	Experimental structures or conformer ensembles, protonation state, geometry source, and conformational metadata	Geometry-aware pretraining or downstream structural refinement	Flexible macrocycles and glycosylated compounds may have uncertain biologically relevant conformations	Conformer-sensitivity analysis and evaluation against suitable experimental geometry where available	A single generated conformer is treated as biologically definitive	Conditional geometric-representation layer
Bioactivity evidence and assay context	Are observed activities separated by endpoint, assay, system, conditions,	Quantitative results, units, relation operators, assay format, target construct,	Context-conditioned fine-tuning, evidence retrieval, and bioactivity prioritization	Sparse, selectively reported, heterogeneous, or conflicting natural product measurements	Assay-stratified evaluation, replication checks, counter-screening, and external testing	Incompatible assays are merged and predictions are mistaken for observed pharmacology	Observed-evidence and bioactivity-context layer

	and evidence status?	species, controls, and publication provenance					
Target annotations	Is each target link identified as measured, curated, inferred, or predicted?	Target identifiers, evidence type, assay context, confidence, and source	Joint compound–target representation and task-specific fine-tuning	Many phytochemical target assignments arise from indirect or computational evidence	Target-family holdout testing and independent interaction confirmation	Predicted associations are interpreted as confirmed engagement	Target-evidence bridge
Evidence provenance and data quality	Can each structure, label, annotation, and transformation be traced and audited?	Source identifiers, database version, publication, extraction method, curator action, and conflict status	Provenance-conditioned weighting, traceable retrieval, and data-quality filtering	Aggregation can replicate incorrect, duplicated, or weakly supported records	Record-level lineage audit and versioned curation	Unverifiable evidence is blended with curated observations	Mandatory governance layer
Uncertainty and bias	Does the representation distinguish model uncertainty, data uncertainty, assay variability, and coverage limitations?	Prediction distributions, domain-distance measures, label variance, scaffold and taxonomic coverage, and subgroup metadata	Calibration, abstention, coverage monitoring, and bias diagnostics	Rare scaffolds, taxa, targets, and assays may lie outside the model's support	Calibration, selective-prediction, scaffold, taxon, target, and temporal external evaluation	Overconfident outputs reflect sampling bias or unsupported extrapolation	Confidence, applicability-domain, and abstention layer

Target context and therapeutic potential

Target-context encoding should begin by separating target identity from target evidence. A phytochemical may be associated with a protein because of a direct binding assay, a functional experiment, curated literature, similarity-based inference, network analysis, or a model prediction, and these relationships should not be represented as equivalent. Transfer learning has been applied to natural product target prediction by using broader chemical information before adapting the model to natural product-specific target labels [16]. This approach supports domain adaptation, but its output remains a ranked target hypothesis whose credibility depends on the similarity between pretraining and target domains, the quality of natural product labels, and the independence of evaluation data. Comparative work on natural product target prediction further indicates that algorithm choice, molecular features, activity curation, and external validation can materially affect apparent performance [17].

Target confidence should therefore combine, but not conflate, model confidence and evidentiary confidence. Model confidence concerns the stability or calibrated probability of a prediction, whereas evidentiary confidence concerns the quality, directness, reproducibility, and contextual relevance of the observations supporting a target association. Machine-learning assessment of natural-compound bioactivity has shown how computational predictions can be linked to literature evaluation and selected experimental testing, providing a useful example of prediction being treated as a hypothesis requiring confirmation rather than as a validated pharmacological conclusion [18]. Sequence-based affinity models likewise demonstrate that molecular strings and protein sequences can be encoded jointly to predict continuous drug–target relationships [19]. Such outputs may assist prioritization, but they do not independently establish physical binding, functional modulation,

selectivity, cellular target engagement, or activity under physiologically relevant exposure conditions.

A richer target-context layer should represent compound–target interactions, target confidence, pathway relationships, disease or phenotype context, and mechanism hypotheses as distinguishable, provenance-linked relations. Multi-objective neural architectures can jointly predict compound–protein affinity and interaction-related patterns, illustrating how structural and target information may be represented together [20]. However, predicted contacts or affinities remain computational outputs rather than experimental interaction maps. Heterogeneous biomedical knowledge graphs further demonstrate how compounds can be connected to genes, pathways, diseases, phenotypes, and other evidence types for prioritization [21]. In the proposed framework, these contextual edges are typed by source and evidence level, enabling a mechanism hypothesis to be assembled without presenting network proximity as causal proof. ADMET and toxicity information should be represented through endpoint-specific outputs whose interpretation remains conditioned by assay domain, data coverage, and external testing requirements [22]. **Figure 1** illustrates how molecular foundation models can connect phytochemical structure, bioactivity evidence, target context, pathway context, safety evidence, uncertainty, and therapeutic-potential boundaries.

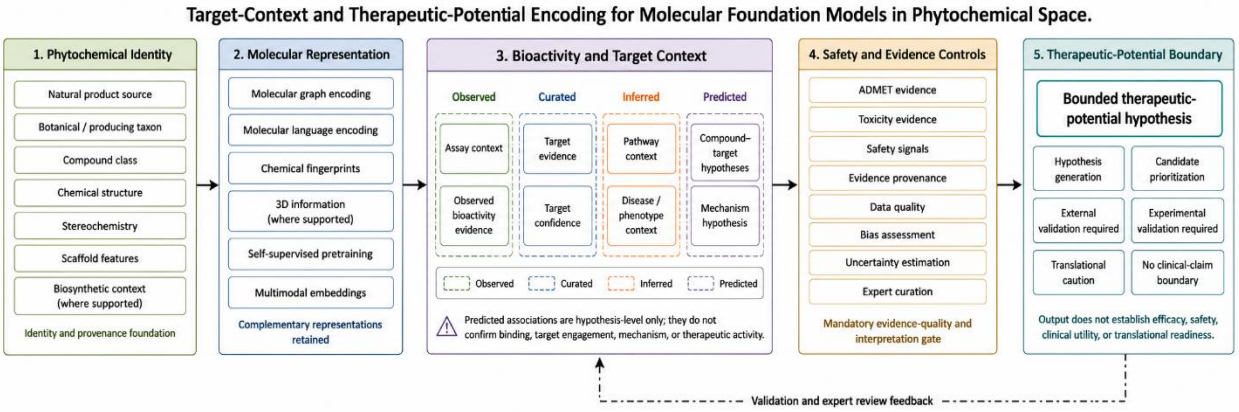

Figure 1. Target-Context and Therapeutic-Potential Encoding for Molecular Foundation Models in Phytochemical Space

Alt-text description

A conceptual encoding diagram showing phytochemical structure and bioactivity evidence linked to target context, pathway context, safety evidence, uncertainty, validation needs, and therapeutic-potential boundaries.

Therapeutic potential should be treated as a bounded hypothesis produced from a structured evidence profile rather than as a latent property that a foundation model can directly reveal. A research-prioritization output may integrate phytochemical identity, structural novelty, assay-qualified bioactivity, target evidence, pathway relevance, disease context, exposure-related considerations, ADMET evidence, toxicity signals, provenance, uncertainty, and validation gaps. The framework should nevertheless prevent a high-ranking output from being translated into claims of efficacy, safety, mechanistic validation, clinical usefulness, or translational readiness. External evaluation is required to determine whether the representation transfers beyond closely related compounds and familiar targets, while experimental confirmation is required for each biological proposition being advanced. Expert review remains necessary to judge chemical identity, assay relevance, target plausibility, safety interpretation, and whether the proposed evidence chain supports further investigation.

Foundation model design requirements

A phytochemical foundation model should be evaluated as a representation system whose usefulness depends on its data foundation, downstream task, and decision context. Learned embeddings should therefore be compared with chemically meaningful fixed descriptors, fingerprints, and task-specific models rather than being assumed superior because they originate from large-scale pretraining. Comparative evaluations of learned molecular representations show that model performance varies across endpoints, datasets, representations, and split strategies, supporting the inclusion of strong conventional baselines in every downstream assessment [23]. Large-scale comparisons of drug-target prediction methods similarly indicate that architecture cannot be interpreted independently of training scale, assay-label construction, chemical features, and competing methods [24]. The proposed design therefore requires natural product-aware

corpus documentation, stereochemistry-sensitive inputs, assay-qualified labels, baseline comparison, and explicit applicability-domain analysis.

Evaluation splits and data curation must also be designed to prevent structural leakage and shortcut learning. Natural product collections contain families of closely related analogues, repeated scaffold series, duplicated records, and compounds reported across multiple databases or studies. When structurally similar molecules occur in both training and evaluation sets, apparent transfer performance may largely reflect memorization of local chemical neighbourhoods rather than generalization to new phytochemical regions. Ligand-based benchmark analyses have demonstrated that conventional classification settings can reward memorization when chemical similarity is not adequately controlled [25]. Chemical dataset investigations further show that hidden biases can materially influence virtual-screening evaluations and should be measured before performance is interpreted [26]. Scaffold-, source-, taxon-, target-, assay-, and temporal separation should consequently be considered during external evaluation, with duplicate resolution and provenance checks completed before model training.

Uncertainty should be represented at several levels. Model uncertainty concerns instability or limited knowledge in the fitted predictor; data uncertainty arises from noisy, conflicting, censored, or incompletely reported measurements; assay uncertainty reflects experimental variability and context dependence; and evidentiary uncertainty concerns the strength of support for a target, pathway, safety, or therapeutic interpretation. Evaluations of scalable uncertainty methods for molecular-property prediction show that uncertainty estimates differ in calibration and usefulness across modeling settings [27]. A phytochemical foundation model should therefore report calibrated predictive uncertainty together with domain-distance indicators and should be capable of abstaining when a structure, target, assay, or contextual combination lies outside supported regions. Confidence values should not be presented as substitutes for biological evidence, and a narrow numerical interval should not be interpreted as confirmation of mechanism, safety, or therapeutic usefulness.

Interpretability, expert curation, and transparent reporting complete the design requirements. Explanatory methods may identify influential atoms, fragments, tokens, graph regions, target features, or contextual relationships, but these explanations remain model-dependent interpretations rather than causal pharmacological evidence. Explainable-AI scholarship in drug discovery emphasizes both the potential utility and the limitations of interpreting complex molecular predictions [28]. In the proposed framework, explanations are intended to support chemical review, identify possible data artifacts, reveal reliance on implausible features, and guide experimental design. Expert curators should be able to inspect source records, stereochemical assignments, assay semantics, target evidence, safety annotations, and unresolved conflicts. Model documentation should disclose the intended use, unsupported uses, corpus composition, preprocessing, representation choices, training objectives, splits, baselines, evaluation metrics, uncertainty methods, known biases, validation status, and claim boundaries. **Table 2** maps design requirements for molecular foundation models intended to encode natural product structure, bioactivity, target context, and therapeutic-potential boundaries.

Table 2. Foundation Model Design Requirements for Phytochemical Space: Natural Product-Aware Pretraining, Stereochemistry-Aware Representation, Bioactivity-Context Encoding, Target-Context Encoding, Multimodal Integration, Evidence Provenance, Uncertainty, Bias Assessment, Safety Awareness, and Validation Boundaries

Design requirement	Purpose	Required data or model feature	Evaluation need	Interpretability need	Failure mode if absent	Translational boundary	Framework output
Natural product-aware pretraining	Ensure that rare scaffolds, complex architectures, source metadata, and natural product classes are represented	Curated phytochemical structures, source annotations, scaffold inventories, compound classes, deduplication, and corpus documentation	Coverage analysis across scaffolds, classes, taxa, stereochemical states, and source databases	Disclose which phytochemical regions are represented or underrepresented	Generic chemical patterns dominate and rare natural product features remain poorly encoded	Broad pretraining does not establish task-specific biological relevance	Natural product-domain coverage profile

Stereochemistry-aware representation	Preserve distinctions that may influence recognition, conformation, and activity	Stereo-aware graphs, isomeric strings, geometric descriptors where supported, and undefined-stereochemistry flags	Stereoisomer challenge sets and evaluation of incomplete or conflicting stereochemical records	Display which stereochemical assignments informed an output	Stereoisomers collapse into indistinguishable or misleading embeddings	Stereo-aware encoding does not establish a biologically active configuration	Stereo-qualified molecular representation
Bioactivity-context encoding	Prevent incompatible activity measurements from becoming interchangeable labels	Endpoint identity, assay format, units, species, target construct, conditions, controls, and evidence provenance	Assay-stratified evaluation, conflict analysis, and external testing	Expose the assay records underlying each prediction	Heterogeneous labels create artificial structure–activity patterns	Predicted bioactivity does not establish pharmacological activity	Assay-conditioned bioactivity hypothesis
Target-context encoding	Distinguish target identity, target evidence, confidence, and biological setting	Target identifiers, protein features, assay evidence, confidence categories, pathway links, and disease context	Target-family holdout, external-target testing, and evidence-tier analysis	Trace each target hypothesis to supporting observations or predictions	Model similarity is interpreted as target engagement or mechanism	Target prioritization requires biochemical or functional confirmation	Ranked target-context hypothesis
Multimodal data integration	Link structure with source, spectra, genomics, assays, targets, pathways, literature, and safety data	Modality-specific encoders, alignment objectives, missing-modality handling, and provenance-preserving fusion	Ablation, cross-modal retrieval, missing-data robustness, and modality-dominance analysis	Retain modality-level contributions and source traceability	One abundant modality overwhelms sparse but important evidence	Integrated embeddings do not convert weak evidence into validated evidence	Decomposable multimodal representation
Evidence provenance tracking	Preserve the origin, transformation history, and evidentiary status of every input and relation	Record identifiers, database versions, publications, extraction methods, curator actions, and evidence types	Record-lineage audit and reproducibility checks	Permit inspection of source records and transformations	Unsupported or duplicated claims become indistinguishable from curated observations	Provenance improves auditability but does not ensure correctness	Traceable evidence bundle
Uncertainty estimation	Prevent unsupported confidence in underrepresented compounds,	Ensembles or calibrated uncertainty methods, domain-distance measures,	Calibration, selective prediction, coverage, and out-of-domain testing	Separate predictive confidence from evidence strength	High-confidence outputs are produced outside the training domain	Uncertainty assists prioritization but does not validate an output	Calibrated prediction and abstention status

	targets, assays, or contexts	label variance, and abstention rules					
Bias and coverage assessment	Identify sampling, structural, taxonomic, assay, target, and publication biases	Coverage metadata, subgroup labels, leakage controls, scaffold clusters, and source statistics	Subgroup, scaffold, taxon, target, assay, temporal, and source-held-out evaluation	Report known blind spots and failure patterns	Performance reflects curation shortcuts or overrepresented chemical families	Bias assessment limits interpretation but cannot remove every bias	Applicability and bias report
Data leakage prevention	Ensure evaluation measures transfer rather than duplication or close-analogue memorization	Structure standardization, duplicate resolution, scaffold-aware splitting, provenance matching, and temporal separation	Similarity audit between training and evaluation records	Reveal overlap and nearest-neighbour relationships	Benchmark results overstate generalization	Internal benchmark success is insufficient for translational interpretation	Leakage-controlled evaluation profile
Safety-aware representation	Ensure candidate prioritization includes endpoint-specific hazard information	ADMET records, toxicity evidence, reactive motifs, off-target data, exposure context, and uncertainty	Endpoint-specific external evaluation and conflict analysis	Decompose safety-related outputs by endpoint, source, and evidence type	Candidates are prioritized without visibility of possible hazards	Predicted absence of risk does not establish safety	Safety-aware research-priority profile
External validation	Test transfer beyond training data, familiar scaffolds, and common targets	Frozen model, independent datasets, prespecified endpoints, and domain-distance measures	Independent scaffold, target, source, assay, or temporal testing	Show how evaluation data differ from development data	Internal validation is presented as evidence of broad generalization	External performance remains task- and domain-specific	External-generalization profile
Experimental validation linkage	Convert predictions into explicit, testable research hypotheses	Output-to-assay mapping, validation priorities, controls, and evidence-gap annotations	Prospective testing with replication and orthogonal methods	State which predicted relation requires which experiment	Predictions accumulate without empirical correction	Experimental confirmation remains limited to the tested endpoint and context	Validation-ready hypothesis package
Expert curation	Resolve chemical identity, taxonomy, stereochemistry, assay semantics, and evidentiary conflicts	Curator interfaces, versioned annotations, conflict states, and decision logs	Inter-curator agreement and audit sampling	Record curator decisions and unresolved uncertainty	Automated errors become embedded and propagated	Expert review supports research decisions but is not clinical validation	Curated evidence state

Transparent reporting	Enable reproducibility, comparison, governance, and responsible interpretation	Data sheets, model cards, preprocessing records, objectives, splits, baselines, metrics, and intended-use statements	Independent audit or reproduction where feasible	Disclose design choices, limitations, and unsupported uses	Results cannot be critically interpreted or reproduced	Transparent reporting does not establish clinical readiness	Foundation-model reporting package
Translational boundary control	Prevent research outputs from becoming unsupported therapeutic or clinical claims	Evidence grades, claim templates, abstention rules, mandatory caveats, and human review	Claim-compliance audit and scenario-based review	Display validation status and prohibited interpretations	A prioritization score is described as efficacy, safety, or readiness	Outputs support research prioritization only	Bounded therapeutic-potential hypothesis

Proposed framework

The proposed Molecular Foundation Model Framework for Phytochemical Space begins with a natural product-aware data foundation. This foundation contains standardized molecular identities, stereochemical assignments, scaffold features, compound classes, source organisms, taxonomic information, literature provenance, biosynthetic context where supported, bioactivity records, assay metadata, target annotations, pathway relationships, ADMET observations, toxicity evidence, and explicit uncertainty fields. Machine learning and genomics scholarship in natural product science indicates that chemical structure can be productively connected with biosynthetic and genomic information, while also emphasizing the heterogeneity and incompleteness of the underlying evidence [29]. The framework therefore treats biosynthetic and genomic context as evidence-qualified modalities: they are incorporated where supported, omitted where unavailable, and never inferred as established facts solely because a structurally related compound has a known biosynthetic origin.

The second stage is representation learning. Separate encoders may process molecular graphs, molecular strings, stereochemical attributes, fixed fingerprints, geometric information where supported, source metadata, bioactivity records, target information, pathway relationships, and text-derived evidence. Self-supervised objectives can be used to learn recurring structural and contextual patterns before fine-tuning, while transfer learning can adapt broad representations to natural product-specific tasks. Natural product cheminformatics demonstrates the value of combining structure normalization, similarity analysis, chemical-space exploration, bioactivity modeling, and target-related inference, but also shows that these methods answer different questions and possess different limitations [30]. The proposed framework therefore avoids a single undifferentiated embedding as the only output. Instead, it retains a shared representation for transfer alongside modality-specific records that remain available for interpretation and audit.

The third stage is context encoding. Observed bioactivity is linked to assay conditions, endpoint definitions, units, biological systems, evidence type, and provenance. Target relations are typed as measured, curated, inferred, or predicted, while pathway and disease-context links carry separate confidence and evidence fields. Knowledge-graph and multimodal approaches in natural product science offer a means of organizing chemical structures, spectra, organisms, literature, biological activities, and contextual relationships without reducing every relation to a structure-only prediction [31]. Within the framework, contextual graphs do not function as proof-generating systems. They organize evidence, reveal missing links, support retrieval, and enable testable compound–target–pathway–phenotype hypotheses. Contradictory records are preserved as conflicts rather than resolved automatically through averaging or majority voting.

The fourth stage introduces safeguards and research-output generation. Uncertainty estimation, bias and coverage assessment, interpretability, data-leakage prevention, provenance tracking, expert curation, and external-validation planning are applied before candidate prioritization. Research outputs may include phytochemical-space maps, scaffold neighbourhoods, target-context hypotheses, assay-specific bioactivity priorities, ADMET-aware filters, safety-signal alerts, evidence-gap summaries, and proposed experimental sequences. Critical assessments of AI in drug discovery argue that model value should be judged by its contribution to a defined scientific decision rather than by architectural

novelty or isolated benchmark performance [32]. Accordingly, the framework does not define a high model score as therapeutic potential. It defines a bounded therapeutic-potential hypothesis as an evidence-weighted research proposition that remains conditional on external testing, experimental confirmation, exposure relevance, safety assessment, and expert interpretation.

The fifth stage establishes governance and translational boundaries. Chemical and biological data differ in semantics, measurement processes, reliability, scale, and experimental context; combining them without preserving these distinctions risks producing internally coherent but scientifically misleading outputs [33]. The framework therefore requires explicit evidence grades, claim controls, versioned curation, model and data documentation, output-specific validation plans, and abstention when support is inadequate. **Table 3** presents the proposed molecular foundation model framework for phytochemical space.

Table 3. Proposed Molecular Foundation Model Framework for Phytochemical Space: Phytochemical Identity, Structure Encoding, Bioactivity Encoding, Target Context, Pathway Context, ADMET and Safety Evidence, Evidence Provenance, Uncertainty, Validation Planning, and Therapeutic-Potential Boundaries

Framework stage	Core purpose	Required input	Model representation logic	Potential output	Validation requirement	Failure risk	Decision boundary	Research value
1. Phytochemical identity and source reconciliation	Establish what the compound is and where the record originated	Standardized structure, identifiers, stereochemistry, compound class, source organism, taxon, material, publication, and database version	Link molecular identity to source- and record-level provenance while preserving conflicts	Curated phytochemical identity record	Expert structure and taxonomy review, duplicate audit, and source verification	Incorrect identity or source association contaminates later representations	Identity reconciliation does not establish biological relevance	Reliable entry point for chemical-space mapping and evidence integration
2. Natural product-aware data foundation	Define the pretraining and contextual corpus	Curated structures, scaffolds, classes, source metadata, bioactivity, targets, pathways, safety records, and evidence grades	Construct coverage-aware, deduplicated, modality-labelled training resources	Corpus coverage and data-quality profile	Scaffold, class, taxon, assay, target, and source coverage audit	Generic or overrepresented chemical regions dominate learning	Corpus size does not imply pharmacological validity	Domain-aware pretraining foundation
3. Structure and stereochemistry encoding	Represent molecular connectivity and configurational information	Molecular graphs, molecular strings, fingerprints, scaffold annotations, stereochemistry, and geometry where supported	Learn complementary structural embeddings while retaining fixed baselines and stereo flags	Structure-aware phytochemical embedding	Scaffold-held-out, stereoisomer, notation-robustness, and geometry-sensitivity testing	Rare scaffolds or stereochemical distinctions are lost	Structural similarity does not establish shared activity	Improved organization, retrieval, and transfer across phytochemical space

4. Self-supervised and multimodal representation learning	Learn transferable patterns across available modalities	Structures, source metadata, text, spectra, genomic context, assays, targets, and pathways	Use modality-specific encoders, alignment objectives, missing-modality handling, and provenance-preserving fusion	Shared and modality-specific representations	Ablation, retrieval, alignment, missing-data, and modality-dominance evaluation	One modality overwhelms others or creates spurious alignment	Cross-modal association does not establish biological causality	Transferable representation and integrated evidence retrieval
5. Bioactivity and assay-context encoding	Separate measured activity from model inference and preserve experimental conditions	Quantitative endpoints, units, relation operators, assay format, species, target construct, controls, and evidence source	Condition activity representations on assay context and evidence type	Assay-specific bioactivity hypothesis or evidence summary	Assay-stratified external evaluation, replication, and counter-screening	Incompatible activities are merged or predictions become pseudo-observations	Bioactivity prediction is not validated pharmacology	Prioritization of relevant assays and compounds
6. Target and pathway-context encoding	Connect compounds with targets, pathways, diseases, and phenotypes at explicit evidence strengths	Target identifiers, interaction evidence, protein features, pathway links, disease context, and confidence categories	Create typed, provenance-linked relations with distinct observed and predicted states	Ranked target or mechanism hypothesis	Independent binding, functional, pathway, or perturbational confirmation	Predicted relationships are described as target engagement or mechanism	Target and pathway context support hypotheses only	Mechanistically organized experimental prioritization
7. ADMET and safety representation	Incorporate endpoint-specific exposure and hazard information	ADME observations, toxicity data, structural alerts, off-target evidence, species, exposure, and provenance	Maintain separate endpoint models and integrated safety summaries with uncertainty	ADMET-aware filter or safety-signal priority	Endpoint-specific external and experimental testing	No predicted alert is interpreted as safety, or an alert as confirmed toxicity	Computational outputs cannot establish safety	Early visibility of possible development and hazard concerns
8. Provenance, quality, bias, and uncertainty control	Make confidence and applicability limitations explicit	Record lineage, quality flags, conflict states, coverage	Weight, filter, abstain, or flag outputs based on evidence	Evidence-confidence profile and abstention status	Calibration, subgroup evaluation, audit, and independe	Overconfident predictions conceal weak data or out-of-domain use	Confidence is not equivalent to evidentiary confirmation	More defensible selection and clearer evidence gaps

ISSN (Online) 2249-6084 (Print) 2250-1029

		metadata, uncertainty estimates, and domain distance	and support		nt data testing			
9. Expert curation and validation planning	Translate outputs into reviewable, testable hypotheses	Model explanations, source records, evidence graphs, conflicts, uncertainty, and candidate-specific gaps	Human review resolves errors, records uncertainty, and assigns validation actions	Validation-ready research hypothesis	Curator audit, experimental controls, replication, and orthogonal methods	Automated errors or plausible narratives guide inappropriate experiments	Expert approval does not replace experimental evidence	Efficient design of targeted follow-up studies
10. Research output generation and therapeutic-potential boundary	Support prioritization while preventing unsupported therapeutic claims	Integrated evidence profile, target context, safety information, uncertainty, and validation status	Produce bounded outputs with claim templates, caveats, and prohibited interpretations	Research-priority ranking or therapeutic-potential hypothesis	External evaluation, experimental confirmation, and downstream translational studies	Model ranking is described as efficacy, safety, clinical utility, or readiness	No autonomous therapeutic or clinical claim	Responsible prioritization of compounds, targets, assays, and validation steps

Figure 2 presents the proposed molecular foundation model framework for phytochemical space, linking natural product-aware pretraining, structure encoding, bioactivity encoding, target context, safety evidence, uncertainty, and validation planning.

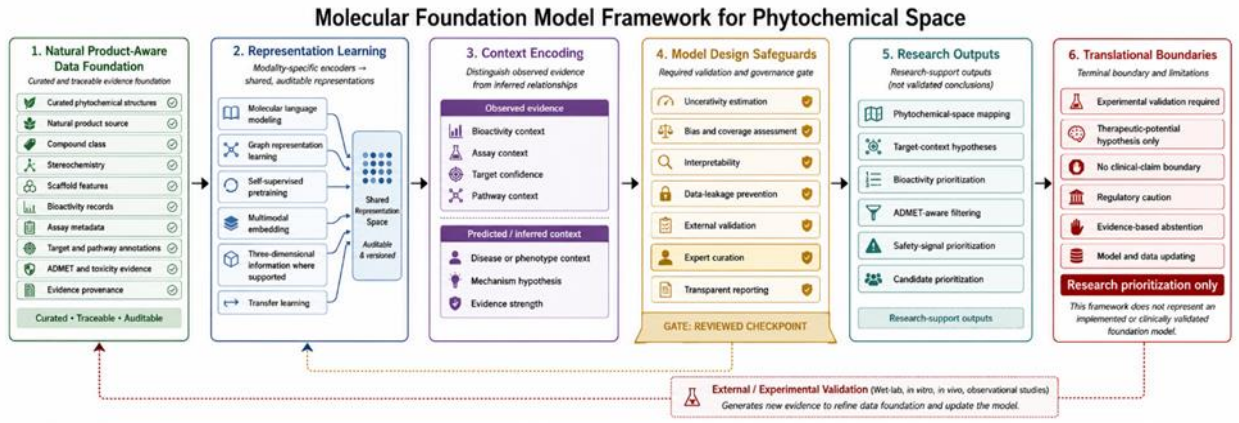

Figure 2. Molecular Foundation Model Framework for Phytochemical Space

Alt-text description

A conceptual framework diagram showing natural product-aware inputs flowing into molecular oundation model pretraining, structure encoding, bioactivity encoding, target-context representation, safety-aware filtering, uncertainty assessment, validation planning, and translational boundaries.

Translational implications

For natural product informatics, the framework shifts database design from structure accumulation toward evidence-preserving representation. Phytochemical resources intended for foundation-model use should support stable chemical identities, stereochemistry,

taxonomic normalization, source provenance, assay semantics, target-evidence categories, safety endpoints, curation histories, uncertainty fields, and versioned conflict resolution. Machine-learning best practices in chemistry emphasize transparent preparation, meaningful evaluation, reproducibility, and disclosure of model limitations [34]. Applied to phytochemical databases, these principles imply that model-ready records should remain traceable to their original evidence and that data transformations should be documented sufficiently for independent audit. Database growth should not be judged solely by record volume; diversity, completeness, quality, and relevance to intended downstream tasks are equally important.

For AI-enabled natural product discovery, molecular foundation models may support chemical-space organization, analogue retrieval, target-hypothesis generation, assay prioritization, ADMET-aware triage, safety-signal identification, and selection of informative experiments. Their role should be integrated with medicinal chemistry, natural product chemistry, pharmacology, toxicology, bioinformatics, and disease-domain expertise. Expert perspectives on AI-assisted drug design stress that computational systems function within multidisciplinary and iterative experimental workflows rather than replacing them [35]. Accordingly, a phytochemical model should provide evidence-linked outputs that researchers can interrogate, challenge, and revise. Candidate prioritization should state why a compound was ranked, which evidence is observed or predicted, where the model is uncertain, what relevant information is absent, and which experiment would most directly reduce the uncertainty.

Responsible therapeutic-potential interpretation requires a strict no-clinical-claim boundary. A target prediction cannot establish mechanism; a pathway association cannot establish disease modification; an activity estimate cannot establish useful exposure; and favourable computational ADMET or toxicity outputs cannot establish safety. Even experimentally confirmed activity remains specific to the tested assay, biological system, dose, and context. Translational progression requires increasingly stringent evidence spanning chemical identity, reproducible pharmacology, target engagement, mechanism, selectivity, exposure, metabolism, toxicity, efficacy models, formulation, manufacturing, and eventually appropriately governed human evaluation. The proposed framework therefore positions molecular foundation models as tools for structured hypothesis generation and research prioritization. It does not position them as autonomous systems for therapeutic validation, regulatory interpretation, or clinical decision-making.

CONCLUSION

This article proposes a molecular foundation model framework designed around the distinctive informational structure of phytochemical space. The framework integrates phytochemical identity, natural product source, chemical structure, stereochemistry, scaffold features, molecular graphs, molecular language, three-dimensional information where supported, bioactivity and assay context, target and pathway relationships, ADMET and toxicity evidence, provenance, data quality, uncertainty, bias assessment, expert curation, external validation, experimental confirmation, and explicit translational boundaries. Its central contribution is to separate representation capacity from evidentiary status: foundation models may improve chemical-space organization, transfer learning, contextual retrieval, hypothesis generation, and research prioritization, but their embeddings and predictions do not independently establish bioactivity, target engagement, mechanism, safety, efficacy, clinical utility, or translational readiness. Progress in phytochemical foundation modeling will therefore depend not only on model scale or architectural sophistication, but also on natural product-aware data curation, stereochemistry-sensitive representation, assay-qualified labels, provenance-preserving multimodal integration, calibrated uncertainty, leakage-resistant evaluation, transparent reporting, and sustained expert and experimental oversight.

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] 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

[3] 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

[4] 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

[5] Sorokina M, Steinbeck C. Review on natural products databases: Where to find data in 2020. J Cheminform. 2020;12(1):20. doi:10.1186/s13321-020-00424-9

[6] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminform. 2021;13(1):2. doi:10.1186/s13321-020-00478-9

[7] Boldini D, Ballabio D, Consonni V, Todeschini R, Grisoni F, Sieber SA. Effectiveness of molecular fingerprints for exploring the chemical space of natural products. J Cheminform. 2024;16(1):35. doi:10.1186/s13321-024-00830-3

[8] Nainala VC, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, et al. COCONUT 2.0: A comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res. 2025;53(D1). doi:10.1093/nar/gkae1063

[9] 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

[10] 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

[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] Zeng X, Xiang H, Yu L, Wang J, Li K, Nussinov R, et al. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. Nat Mach Intell. 2022;4(11):1004-16. doi:10.1038/s42256-022-00557-6

[13] Pinheiro GA, Da Silva JLF, Quiles MG. SMICLR: Contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning. J Chem Inf Model. 2022;62(17):3948-60. doi:10.1021/acs.jcim.2c00521

[14] Kaufman B, Williams EC, Underkoffler C, Pederson R, Mardirossian N, Watson I, et al. COATI: Multimodal contrastive pretraining for representing and traversing chemical space. J Chem Inf Model. 2024;64(4):1145-57. doi:10.1021/acs.jcim.3c01753

[15] Méndez-Lucio O, Nicolaou CA, Earnshaw B. MolE: A foundation model for molecular graphs using disentangled attention. Nat Commun. 2024;15(1):9431. doi:10.1038/s41467-024-53751-y

[16] Qiang B, Lai J, Jin H, Zhang L, Liu Z. Target prediction model for natural products using transfer learning. Int J Mol Sci. 2021;22(9):4632. doi:10.3390/ijms22094632

[17] Liang L, Liu Y, Kang B, Wang R, Sun MY, Wu Q, et al. Large-scale comparison of machine learning algorithms for target prediction of natural products. Brief Bioinform. 2022;23(5). doi:10.1093/bib/bbac359

[18] Periwal V, Bassler S, Andrejev S, Gabrielli N, Patil KR, Typas A, et al. Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs. PLoS Comput Biol. 2022;18(4). doi:10.1371/journal.pcbi.1010029

[19] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: Deep drug-target binding affinity prediction. Bioinformatics. 2018;34(17). doi:10.1093/bioinformatics/bty593

[20] Li S, Wan F, Shu H, Jiang T, Zhao D, Zeng J. MONN: A multi-objective neural network for predicting compound-protein interactions and affinities. Cell Syst. 2020;10(4):308-22.e11. doi:10.1016/j.cels.2020.03.002

[21] Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017;6. doi:10.7554/eLife.26726

[22] Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, et al. ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res. 2021;49(W1). doi:10.1093/nar/gkab255

[23] 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

[24] 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

[25] Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403

[26] Sieg J, Flachsenberg F, Rarey M. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. J Chem Inf Model. 2019;59(3):947-61. doi:10.1021/acs.jcim.8b00712

[27] Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697-717. doi:10.1021/acs.jcim.9b00975

[28] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84. doi:10.1038/s42256-020-00236-4

[29] Prihoda D, Maritz JM, Klempir O, Dzamba D, Woelk CH, Hazuda DJ, et al. The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. Nat Prod Rep. 2021;38(6):1100-8. doi:10.1039/D0NP00055H

[30] Chen Y, Kirchmair J. Cheminformatics in natural product-based drug discovery. Mol Inform. 2020;39(12). doi:10.1002/minf.202000171

[31] Meijer D, Beniddir MA, Coley CW, Mejri YM, Öztürk M, van der Hooft JJJ, et al. Empowering natural product science with AI: Leveraging multimodal data and knowledge graphs. Nat Prod Rep. 2025;42(4):654-62. doi:10.1039/D4NP00008K

[32] Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009

[33] 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 used for AI in drug discovery. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037

[34] Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8. doi:10.1038/s41557-021-00716-z

[35] 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

ISSN (Online) 2249-6084 (Print) 2250-1029