Pharmacophore

ISSN-2229-5402

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

ONE MOLECULAR LANGUAGE FOR CHEMICAL STRUCTURES, BIOLOGICAL SEQUENCES, THREE-DIMENSIONAL CONFORMATIONS, SPECTRA, IMAGES, AND SCIENTIFIC TEXT

Elena Fedorova[1]*, Alexey Volkov[1], Irina Morozova[2], Dmitry Kuznetsov[3]

1. ***Department of Unified Molecular Language, Faculty of Pharmacy, Novosibirsk State University, Novosibirsk, Russia.***
2. ***Department of Chemical Structures and Biological Sequences, Faculty of Pharmaceutical Sciences, Ural Federal University, Yekaterinburg, Russia.***
3. ***Department of Conformations, Spectra, Images, and Text Integration, Faculty of Pharmacy, Kazan Federal University, Kazan, Russia.***

ARTICLE INFO

Received:
03 August 2025
Received in revised form:
30 November 2025
Accepted:
05 December 2025
Available online:
28 December 2025

Keywords: Multimodal molecular representation, Chemical language models, Protein language models, Three-dimensional molecular learning, Molecular spectra, Phenotypic imaging

ABSTRACT

Pharmaceutical knowledge is distributed across chemical graphs and strings, biological sequences, three-dimensional conformations, analytical spectra, cellular images, and scientific text. Although contemporary representation-learning systems can encode several of these sources, their outputs commonly remain divided by incompatible units, acquisition processes, invariances, uncertainty structures, and validation conventions. This fragmentation limits cross-modal retrieval, weakens transfer between molecular and biological contexts, and can encourage the unsupported treatment of correlated representations as interchangeable scientific evidence. This article develops a proposed multimodal molecular-language architecture that distinguishes semantic content that may be shared across modalities from information that must remain modality-specific. The construct combines versioned identity anchors, modality-native representations, a bounded shared semantic layer, private residual channels, provenance records, uncertainty descriptions, contradiction handling, and task-specific evidence gates. Its central premise is that molecular identity, substructure, perturbation, phenotype, function, and experimental context may provide alignment anchors, but alignment does not erase the distinct meanings of coordinates, spectral peaks, image features, sequence patterns, or textual claims. The architecture therefore treats missing modalities, inferred proxies, contextual disagreement, and unresolved contradiction as explicit representational states rather than hidden preprocessing problems. Evaluation is framed as a multidimensional obligation covering retrieval, generation, transfer, native-modality fidelity, uncertainty, contradiction sensitivity, and scientific grounding. The proposal remains conceptual and does not establish empirical superiority, mechanistic validity, therapeutic utility, regulatory acceptability, or deployment readiness. Its principal implication is that progress toward a reusable molecular language requires not merely larger models or unified embeddings, but disciplined preservation of modality-specific evidence and transparent boundaries between learned correspondence and pharmaceutical meaning.

To Cite This Article: Fedorova E, Volkov A, Morozova I, Kuznetsov D. One Molecular Language for Chemical Structures, Biological Sequences, Three-Dimensional Conformations, Spectra, Images, and Scientific Text. Pharmacophore. 2025;16(6):89-98. https://doi.org/10.51847/bWOObpCFdP

Introduction

Pharmaceutical science does not encounter a molecule through one stable representational form. A candidate compound may appear as a graph, a line notation, a stereochemically resolved structure, a conformational ensemble, a mass spectrum, an image-derived perturbational phenotype, a textual assertion, or an entry linked to a protein sequence and biological context. Each representation makes selected properties accessible while suppressing others. A molecular string may encode connectivity efficiently but omit the environmental and conformational states relevant to binding. A coordinate set may express geometry while remaining silent about whether that geometry was experimentally observed, computationally predicted, or selected from a dynamic ensemble. A microscopy image may capture a cellular response without directly identifying its molecular mechanism. Scientific text may connect these sources narratively while mixing observation, inference,

Corresponding Author: Elena Fedorova; Department of Unified Molecular Language, Faculty of Pharmacy, Novosibirsk State University, Novosibirsk, Russia. E-mail: elena.fedorova@nsu.ru

interpretation, and speculation. The unresolved problem is therefore not simply how to concatenate additional data types. It is how to construct a common representational system without converting scientifically different evidence into falsely equivalent tokens.

This problem matters because machine learning in drug discovery is increasingly expected to connect discovery stages, evidence types, and decision contexts. Yet useful pharmaceutical intelligence depends on fit-for-purpose data, interpretable relationships, validation appropriate to the intended task, integration with chemical reasoning, and recognition that generated or predicted outputs remain subordinate to experimental and medicinal-chemistry judgment [1-3]. A system may retrieve structurally similar molecules while missing differences in stereochemistry, assay context, or phenotype. It may generate chemically valid strings while failing to preserve synthetic accessibility or meaningful target engagement. It may align compound descriptions with structures while learning textual co-occurrence rather than experimentally grounded relationships. These failures are not reducible to insufficient model scale. They arise partly because the objects being aligned have different epistemic status and because conventional performance summaries rarely reveal which meanings have been preserved, transformed, or discarded.

Existing molecular and biological language models demonstrate that transferable regularities can be learned within particular representational systems. However, the term *language* can become misleading when syntax learning, embedding similarity, physical structure, biological state, and scientific evidence are treated as if they were interchangeable forms of the same information. Chemical strings are constructed encodings; sequences are inherited or experimentally determined symbolic records; conformations are spatial states or hypotheses; spectra are instrument-conditioned measurements; images are spatial observations shaped by acquisition and preprocessing; and text is a human-produced account of claims and evidence. A unified architecture must therefore support communication among modalities while maintaining distinctions among representation, measurement, prediction, interpretation, and mechanism. Otherwise, apparent unification may amount to semantic compression that hides precisely the differences required for reliable pharmaceutical use.

This article proposes an original multimodal representation architecture for a bounded molecular language spanning chemical structures, biological sequences, three-dimensional conformations, spectra, images, and scientific text. The contribution is architectural rather than empirical: it defines shared semantic obligations, modality-private information, alignment anchors, provenance requirements, uncertainty states, contradiction handling, and task-specific evaluation boundaries. A reusable representation must also be supported by transparent task definitions, suitable datasets, comparable evaluation practices, and explicit restrictions on transfer between contexts that do not share the same evidence-generating process [4]. The central argument is that no single embedding, loss function, or benchmark measure can establish a universal molecular language. Such a language must instead preserve identity and context, align only defensible semantic content, expose what remains unaligned, and evaluate retrieval, generation, transfer, and scientific grounding as separate but connected functions.

Fragmented Molecular Languages across Pharmaceutical Data Modalities

Fragmentation begins within chemical representation itself. Molecular-learning tasks vary in endpoints, dataset composition, class balance, splitting strategy, metric selection, and sensitivity to local chemical neighborhoods. Consequently, a representation that performs well for one property under a random split may be inadequate for a different endpoint, scaffold regime, or extrapolative setting. Fragmentation also arises between two-dimensional chemical encodings and three-dimensional descriptions, because geometry introduces transformation rules and physical relationships that cannot be represented faithfully through arbitrary coordinate handling. Task-dependent benchmark variation and symmetry-dependent geometric obligations therefore represent distinct but connected sources of representational fragmentation [5, 6]. This distinction matters for the proposed molecular language: chemical identity cannot be equated with one canonical string, and structural correspondence cannot be assumed to preserve conformational or interaction-relevant meaning.

Three-dimensional conformations require a modality-native semantics based on positions, distances, angles, local environments, symmetry, and potentially ensembles rather than isolated coordinate sets. Rotating or translating a molecular system should not alter its physical interpretation, while reflection may or may not preserve meaning depending on stereochemical context. Equivariant architectures demonstrate why these transformation properties must be built into representation design rather than learned as incidental statistical regularities [6]. Even then, a three-dimensional encoder does not receive a complete physical state. Protonation, solvation, temperature, binding partners, crystal packing, conformational populations, and measurement or simulation conditions may remain absent. The architecture developed here therefore treats geometry as an indispensable private representational domain connected to molecular identity, but not reducible to connectivity and not automatically sufficient for mechanistic interpretation.

Cellular images expose a different form of molecular meaning. They may record changes in morphology, organelle organization, localization, cell count, texture, or other phenotypic features following chemical or genetic perturbation. Their semantic content depends on segmentation, illumination correction, feature extraction, quality control, normalization, plate design, cell type, dose, time, and imaging protocol. Image-based cell profiling accordingly requires an analytical chain that is distinct from graph or sequence processing before its features can be linked to compounds or targets [7]. A shared molecular language must retain these acquisition and biological-context variables, because visual similarity can reflect common mechanism, convergent downstream effects, technical artifacts, general toxicity, or unrelated morphological responses. Images are therefore not depictions of molecular structure in another visual format; they are contextual measurements of biological state.

Tandem mass spectra further demonstrate why multimodal molecular data cannot be treated as direct translations of one another. Peaks arise from ionization, fragmentation, instrument settings, collision conditions, molecular formula, adduct formation, and structural features. Computational annotation can rank candidate formulas and structures, but the spectrum remains measurement-conditioned evidence rather than a direct molecular string [8]. Isomers may produce similar fragments, relevant peaks may be absent, and library coverage may constrain identification. Across structures, geometries, images, and spectra, fragmentation is thus produced by different representational units, generation processes, invariances, noise sources, and evidentiary meanings. The proposed architecture responds by linking modalities through explicit identity, context, and alignment anchors while prohibiting unsupported equivalence. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop fragmented molecular languages across pharmaceutical data modalities within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Fragmented molecular languages across pharmaceutical data modalities

Construct or evidence domain	Core question	Scientific basis	Role in the article	Failure or overclaiming risk	Boundary statement
Chemical graphs and molecular strings	Which aspects of molecular identity, connectivity, stereochemistry, and composition are preserved by the encoding?	Graphs and line notations provide constructed symbolic descriptions whose utility depends on task, tokenization, canonicalization, and validation regime.	Establishes the chemical-identity and connectivity layer from which cross-modal links may be indexed.	Treating one notation or embedding as a complete description of chemical behavior.	A valid chemical encoding does not establish conformation, activity, synthesizability, safety, or therapeutic value.
Three-dimensional conformations	Which geometric relations and physical symmetries must remain invariant or equivariant?	Coordinates express spatial organization but depend on conformer generation, structural source, transformation rules, and environmental assumptions.	Supplies geometry-specific information that cannot be safely compressed into two-dimensional connectivity alone.	Presenting one predicted or selected coordinate set as the complete physical state of a molecule.	A conformation is a contextual structural hypothesis or observation, not a full account of molecular dynamics or biological function.
Biological sequences	Which evolutionary, structural, and functional regularities are encoded, and at what biological level?	Protein and nucleic-acid sequences are symbolic biological records whose meaning depends on variation, evolutionary context, organization, and molecular environment.	Provides a biological identity and function-oriented modality that may align with structures, targets, and textual evidence.	Assuming that sequence similarity guarantees equivalent structure, function, binding, or phenotype.	Sequence correspondence supports hypotheses but does not alone establish functional or pharmacological equivalence.
Analytical spectra	What structural evidence is contained in the measurement, and how does acquisition context shape it?	Spectral peaks are produced by instrument- and condition-dependent physical measurements and require probabilistic interpretation.	Connects measured chemical evidence to candidate identity and structural hypotheses.	Treating a high-ranking spectral match as definitive molecular identification.	Spectral correspondence is evidence for identity; confirmation may require orthogonal measurements and contextual verification.
Cellular and molecular images	Which phenotypic or spatial features are biological, and which arise from acquisition or preprocessing?	Images capture contextual observations shaped by sample preparation, segmentation, normalization, batch effects, cell state, dose, and time.	Represents downstream biological response and spatial organization that may complement structure and sequence.	Inferring a unique mechanism from morphological similarity or visual model confidence.	Phenotypic similarity does not by itself establish a common target, pathway, mechanism, or causal process.
Scientific text	Which statements are observations, interpretations, hypotheses, or reproduced claims, and what	Text encodes human scientific communication, including evidence, uncertainty, negation, terminology, and reporting bias.	Provides explanatory and relational context while linking entities, experiments, and claims.	Equating linguistic fluency, co-occurrence, or retrieval relevance with experimental truth.	Textual alignment must remain attached to source provenance and assertion status; language is not experimental confirmation.

	provenance supports them?				
Identity and provenance	Are records linked to the same entity, form, sample, perturbation, and evidence-generating context?	Salts, stereoisomers, tautomers, conformers, sequence variants, mixtures, samples, and assay records may share names while differing scientifically.	Prevents accidental fusion and supplies versioned anchors for cross-modal relationships.	Collapsing related but non-equivalent records under a shared identifier.	Entity linkage enables comparison but does not establish biological equivalence or interchangeable evidence.
Cross-modal correspondence	Which semantic relation justifies alignment between two modalities?	Defensible anchors may include identity, substructure, sequence–structure relation, perturbation, target, phenotype, or source-supported assertion.	Defines the limited content that may enter a shared semantic layer.	Maximizing embedding proximity without demonstrating what correspondence has been learned.	Alignment denotes a modeled relationship and must not be reported automatically as mechanism, causality, or scientific confirmation.

Shared and Modality-Specific Semantics

A multimodal molecular language requires a distinction between *shared semantics* and *modality-specific semantics*. Shared semantics are concepts for which a defensible relationship can be established across evidence types, such as molecular identity, substructure, residue position, target association, perturbation identity, experimental condition, phenotype, or a source-linked scientific claim. Modality-specific semantics are properties whose meaning depends on the native representation or acquisition process, including graph topology, reaction direction, geometric orientation, conformational multiplicity, spectral intensity, image texture, sequence context, or textual assertion status. The distinction is functional rather than absolute. A concept may be shareable at one level while retaining private detail at another. For example, a reaction record can link reactants and products across chemical strings and text, but its yield, conditions, uncertainty, and mechanistic interpretation cannot be assumed to reside completely in a shared embedding. Sequence-based reaction modeling illustrates that chemical relationships can be learned through language-like representations, while calibrated confidence and domain coverage remain necessary for interpreting predictions [9].

Chemical language pretraining provides evidence that large molecular-string corpora contain reusable regularities associated with structure and properties. Such representations can support transfer because tokens and contexts repeatedly encode fragments, functional groups, connectivity patterns, and broader molecular distributions. However, the learned semantics remain conditioned on the selected notation, tokenization, corpus composition, preprocessing decisions, and downstream tasks [10]. A shared semantic layer should therefore not be defined as whatever information a large encoder happens to compress. It should be constrained by explicit semantic obligations. Molecular identity links should preserve stereochemical and form-related distinctions; substructure relations should remain traceable to chemical features; and property-oriented transfer should state which endpoint and domain support the relationship. Information that cannot be expressed safely through those obligations should remain in a modality-private channel rather than being forced into a common latent space.

Biological-sequence models similarly show that unsupervised learning can recover patterns associated with evolution, structure, and function. These emergent representations are valuable because sequence corpora contain relationships across residues, domains, families, and selective constraints [11]. Nevertheless, protein-sequence semantics are not interchangeable with small-molecule semantics. A residue token participates in an evolved polymer with positional, structural, and functional dependencies; a molecular token belongs to a constructed chemical notation whose syntax and equivalence rules differ. The proposed language therefore allows sequence and chemical representations to meet through bounded anchors, such as experimentally supported interactions, binding-site relations, perturbational associations, or source-linked target annotations. It does not assume that a single tokenizer, distance metric, or embedding geometry is scientifically appropriate for both. Cross-modal proximity must be interpretable in terms of an identified relation, not merely a favorable contrastive-learning objective.

The relationship between biological sequence and three-dimensional protein structure further clarifies why shared semantics must coexist with modality-private information. Structure prediction can infer spatial organization from sequence and evolutionary context, and confidence estimates can identify regions of differing reliability [12]. Yet the resulting coordinates remain modeled structural information rather than experimentally observed geometry, and they may not capture alternative conformations, disorder, complexes, ligands, post-translational states, or environmental dynamics. In the proposed architecture, sequence and structure can share residue identity, order, domain organization, and hypothesized spatial relations, while geometric details, prediction provenance, confidence, and conformational uncertainty remain explicitly attached to the structural modality. The same principle extends to spectra, images, and text: the shared layer should encode only supported correspondences, whereas native residual channels preserve evidence that would be distorted by universal fusion. One molecular language is therefore not a demand that every modality say the same thing. It is a disciplined system for stating what they can say together, what each can say alone, and how uncertainty qualifies both.

Architecture of the Multimodal Molecular Language

The proposed architecture begins by separating five operations that are often conflated: representation, alignment, translation, fusion, and co-learning. Representation converts native inputs into modality-aware features; alignment identifies defensible correspondences; translation generates or retrieves one modality from another; fusion combines evidence for a defined task; and co-learning transfers information when paired modalities are incomplete or unevenly available. These operations answer different scientific questions and require different validation designs [13]. The architecture therefore rejects a monolithic latent space in which all inputs are compressed indiscriminately. Instead, it uses versioned entity anchors, dedicated encoders, bounded alignment interfaces, modality-private residual channels, and task-specific evidence gates.

At its center is a molecular-identity and provenance layer that distinguishes chemical forms, sequence variants, conformers, measurements, samples, perturbations, and textual records. This layer establishes which records may be compared without declaring them scientifically identical. Paired structure–text learning shows that molecular representations and biomedical language can support cross-modal retrieval and comprehension, but the resulting relationship remains dependent on corpus construction, entity linking, and textual evidence quality [14]. The proposed observatory consequently records whether each relationship is observed, curated, predicted, inferred, or textually asserted. **Figure 1** presents the universal molecular language observatory, showing how the article's key components and boundaries are connected within architecture of the multimodal molecular language.

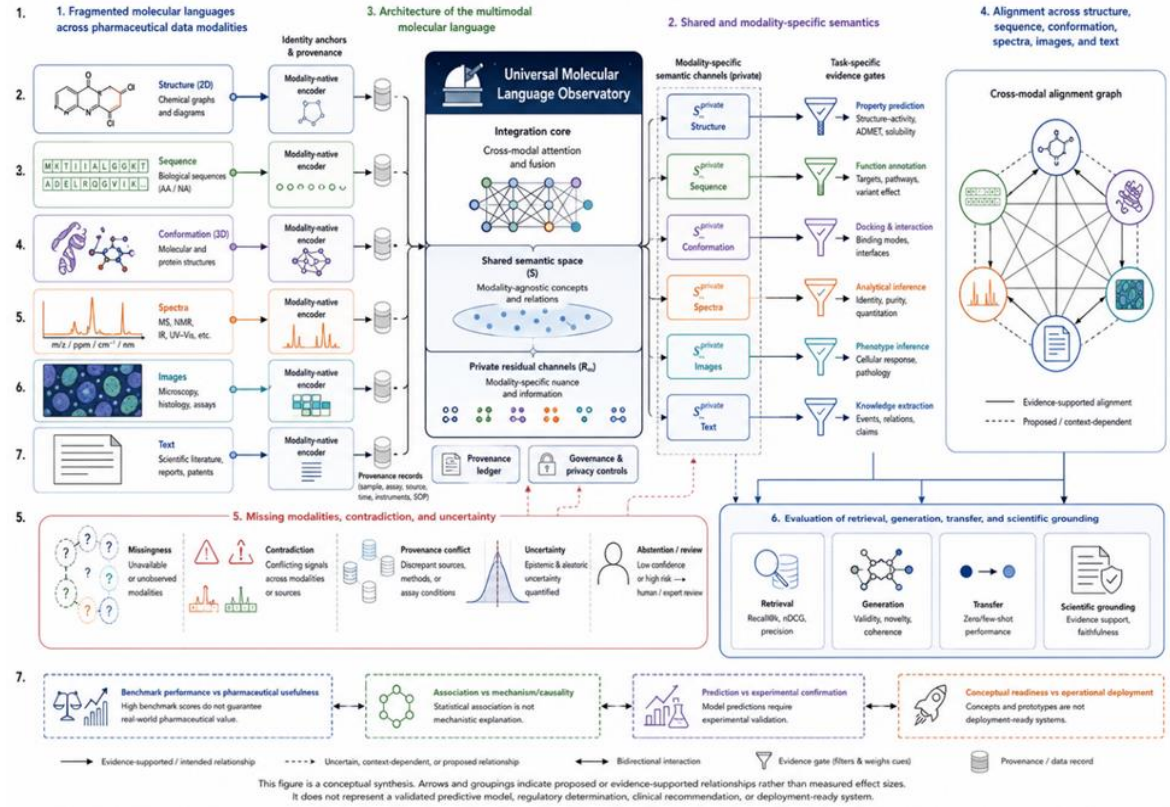

Figure 1. Universal Molecular Language Observatory

The shared semantic core receives only relations supported by identity, structural correspondence, perturbation, target, phenotype, or provenance-aware text. Parallel private channels retain information that should not be forced into this core, including stereochemical detail, conformational alternatives, spectral acquisition conditions, image texture, sequence context, and textual assertion status. Multimodal chemical pretraining demonstrates that representations can be connected across molecular formats and traversed within a shared chemical space [15]. In the proposed architecture, however, shared proximity is always accompanied by modality labels and decoding constraints. This prevents an aligned latent representation from being interpreted as proof that its contributing modalities contain equivalent evidence.

Separate decoders support retrieval, editing, generation, transfer, prediction, and grounded explanation. Each decoder receives only the shared and private information required for its task and is governed by a distinct evidence gate. Structure–text alignment may support bidirectional retrieval or controlled molecular editing, but a text-conditioned output must still be checked for chemical validity, identity preservation, uncertainty, and consistency with the requested constraints [16]. Successful decoding therefore establishes task-level representational competence, not synthesizability, target engagement, safety, therapeutic benefit, or mechanistic truth. The architecture's defining principle is controlled interoperability: modalities can exchange information without surrendering their native epistemic boundaries.

Alignment across Structure, Sequence, Conformation, Spectra, Images, and Text

Alignment begins with an explicit statement of the relation being learned. Two-dimensional molecular depictions can support property and target prediction because visual arrangements encode connectivity, atom types, bonds, and recurring structural patterns [17]. Their semantics nevertheless depend on depiction conventions, resolution, orientation, annotation, and rendering choices. A depiction encoder may align with a graph through atom and bond correspondence, but it should not be treated as equivalent to a conformational representation. Graph–image agreement tests representational consistency; it does not show that either modality contains the spatial, energetic, or environmental information needed for a three-dimensional claim.

Spectral alignment requires a different anchor. Tandem mass spectra are linked to candidate structures through precursor identity, formula constraints, fragmentation patterns, adduct information, and acquisition metadata. Large-scale self-supervised learning can extract reusable structure-relevant representations from spectral collections [18]. However, a spectrum-native embedding remains conditioned by ionization, collision energy, instrumentation, sample composition, and observable fragments. Structure–spectrum alignment should therefore be probabilistic and provenance-aware. The architecture must retain alternative candidates and missing peaks rather than forcing each spectrum into a single deterministic molecular identity.

Molecular depictions, tandem mass spectra, and cell-response profiles provide distinct routes to molecular meaning rather than interchangeable views [17-19]. The same principle governs all six modalities. Structure–sequence alignment should be anchored to residue identity, binding sites, or experimentally supported interactions; structure–conformation alignment should preserve atom correspondence and geometric symmetry; structure–text alignment should distinguish entity description from evidence-bearing claims; and compound–phenotype alignment should retain dose, time, cell state, and assay context. Agreement may strengthen a task-specific hypothesis, whereas disagreement may expose biological adaptation, measurement limitations, representation loss, or incompatible contexts.

Paired perturbational datasets provide a particularly useful substrate because the same chemical or genetic intervention can be associated with morphology and gene-expression profiles. These modalities capture overlapping but non-identical aspects of cellular state [20]. Alignment should therefore be conditional on perturbation identity, biological system, dose, exposure duration, and acquisition process. The proposed architecture represents concordant features in the shared layer while preserving modality-private responses and explicit discordance. It does not demand that transcriptomic and morphological representations converge completely. Their differences may contain the evidence needed to recognize delayed responses, pathway-specific effects, general stress, or technically induced disagreement.

Missing Modalities, Contradiction, and Uncertainty

Complete multimodal records are uncommon. A compound may have a structure and textual description but no spectrum, conformational ensemble, phenotypic image, or matched biological sequence context. Missing-modality learning can transfer information from richer multimodal records to settings where only one input type is available [21]. The proposed architecture permits such transfer through an availability-aware router, but every modality is tagged as observed, absent, or inferred. Reconstructed representations remain proxy modalities rather than measurements. Their use must be restricted when the missingness mechanism differs from training conditions or when a decision depends directly on the unavailable evidence.

Contradiction must also be separated from technical discordance. Imaging profiles can vary because of plate effects, laboratory conditions, preprocessing choices, cell density, instrumentation, or normalization. Comparative evaluations of batch correction show that harmonization methods can alter both technical variation and biologically meaningful structure [22]. The architecture therefore applies provenance and batch diagnostics before cross-modal fusion. Correction is not assumed to reveal a single true representation, and residual disagreement is retained. Where modalities arise from different contexts, their divergence may be scientifically expected rather than evidence of model failure.

Uncertainty is recorded as a structured ledger rather than a single confidence score. The ledger distinguishes input uncertainty, model uncertainty, alignment uncertainty, missingness uncertainty, domain-shift uncertainty, and task-specific decision uncertainty. Comparative studies of molecular prediction show that uncertainty methods can capture different error components and behave differently under distribution shift [23]. The architecture consequently prohibits the conversion of internal confidence into universal scientific certainty. Every output must identify the source of uncertainty, the domain in which it was assessed, and whether calibration has been examined for the relevant task.

No uncertainty estimator can be assumed to perform consistently across molecular datasets, endpoints, architectures, and evaluation measures [24]. Uncertainty must therefore be judged against the decision it is intended to support. Ranking uncertain retrieval results, abstaining from generation, prioritizing experiments, and bounding a property prediction are different uses with different failure costs. The proposed system can defer output, request human assessment, or recommend acquisition of a missing modality when uncertainty or contradiction crosses a task-defined boundary. These actions are architectural safeguards requiring future validation, not established decision rules.

Evaluation of Retrieval, Generation, Transfer, and Scientific Grounding

Evaluation begins by refusing to treat one aggregate score as evidence of a universal molecular language. Molecular representations can change their apparent utility across endpoints, datasets, splitting strategies, data scales, and implementation conditions [25]. Retrieval should therefore be tested with provenance-resolved matched pairs, chemically and biologically meaningful hard negatives, and modality-specific error analysis. Transfer should be examined across endpoints and domains,

including cases in which pretraining harms performance. Reproducibility, identity preservation, private-information retention, and sensitivity to acquisition context are distinct obligations rather than auxiliary diagnostics.

Generation requires separate assessment because syntactic validity and distributional similarity do not establish useful molecular design. Benchmark suites can compare distribution-learning and goal-directed generation, but their tasks and metrics may reward narrow optimization or repeated exploitation of benchmark-specific patterns [26]. The proposed evaluation contract therefore separates representation validity, constraint preservation, diversity, novelty, recoverability of conditioning information, uncertainty, and evidence-supported utility. Generated structures or edits must remain distinguishable from experimentally realized compounds and from candidates whose synthesis, activity, selectivity, exposure, or safety has been established.

Standardized datasets and multiple complementary measures can improve comparability among molecular generators [27]. Nevertheless, validity, uniqueness, novelty, and distributional resemblance remain proxy assessments. They do not demonstrate that generated molecules occupy a useful pharmaceutical design space or satisfy downstream experimental requirements. Evaluation should include failure taxonomies, condition-sensitive tests, negative controls, and expert examination of chemically meaningful cases. Model comparison should report trade-offs rather than collapse them into a single rank, especially when improvements in one measure accompany losses in chemical diversity, constraint satisfaction, or robustness.

Scientific grounding asks whether outputs preserve identity, retrieve source-supported relationships, respect local chemical neighborhoods, and fail transparently when small structural changes produce large activity changes. Activity cliffs show that favorable average prediction can coexist with severe local error around highly similar molecules [28]. Grounding evaluation must therefore test neighborhood consistency, stereochemical sensitivity, provenance, contradiction recognition, and correspondence with experimental evidence. It must also distinguish retrieval relevance from factual support, prediction from mechanism, and plausible generation from pharmaceutical usefulness. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop evaluation of retrieval, generation, transfer, and scientific grounding within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from One Molecular Language for Chemical Structures, Biological Sequences, Three-Dimensional Conformations, Spectra, Images

Component or layer	Inputs	Core function	Expected output	Uncertainty or limitation	Validation requirement
Identity and provenance layer	Versioned chemical identifiers, sequence variants, conformers, samples, perturbations, assay conditions, source records	Determines which records refer to the same or related entities without declaring evidence equivalence	Traceable entity links and relation types	Ambiguous names, mixtures, tautomers, salts, stereoisomers, outdated records	Curated identity challenges, provenance audits, and expert review of ambiguous mappings
Modality-native encoders	Graphs, strings, sequences, coordinates, spectra, images, and text	Preserves syntax, invariances, acquisition context, and native information	Modality-specific representations with retained metadata	Corpus bias, preprocessing loss, encoder misspecification	Native reconstruction, invariance tests, perturbation tests, and modality-specific error analysis
Shared semantic layer	Verified cross-modal pairs and explicit alignment anchors	Represents supported identity, substructure, perturbation, target, phenotype, or claim relations	Comparable representations for bounded cross-modal tasks	False alignment, semantic collapse, shortcut learning	Hard-negative retrieval, residual-information probes, ablation, and external-pair evaluation
Modality-private channels	Geometry, sequence context, spectral conditions, image texture, assertion status, and other non-shared information	Retains evidence that cannot be safely projected into the common layer	Recoverable modality-specific detail and discordance	Private information may be ignored by downstream models	Reconstruction, counterfactual removal, and task-specific utility assessment
Retrieval gate	Query modality, candidate modalities, provenance-resolved correspondence labels	Retrieves scientifically corresponding records across modalities	Ranked candidates with relation type and uncertainty	Dataset leakage, metadata shortcuts, semantically weak matches	External retrieval sets, hard negatives, source verification, and expert error classification
Generation and editing gate	Conditioning text, structures, constraints, shared/private representations	Produces a candidate structure, representation, or	Valid candidate with preserved conditions and explicit uncertainty	Invalid chemistry, constraint loss, mode collapse, benchmark gaming	Chemical checks, condition recovery, diversity tests, synthesis-aware assessment, and

		controlled transformation			experimental confirmation where required
Transfer and prediction gate	Pretrained representations, target endpoint, domain metadata, limited labeled data	Transfers learned information to a specified prediction task	Task-bound prediction with applicability statement	Negative transfer, domain shift, endpoint-specific failure	External and temporal testing, learning curves, ablations, neighborhood analysis, and calibration
Missing-modality router	Availability masks, observed modalities, training-time modality patterns	Selects observed pathways and labels inferred proxy information	Bounded output that distinguishes observation from inference	Proxy modalities may appear plausible but be wrong	Planned missingness studies, real incomplete records, and uncertainty–missingness analysis
Contradiction and uncertainty layer	Conflicting modalities, batch metadata, repeat measurements, model outputs	Classifies technical, contextual, biological, and unresolved disagreement	Uncertainty ledger, abstention, review request, or evidence-acquisition recommendation	No unique ground truth; calibration may drift	Replicate measurements, adjudicated contradiction cases, shift testing, and decision-specific calibration
Scientific-grounding gate	Source-linked claims, local chemical neighborhoods, perturbation evidence, expert knowledge	Tests whether outputs remain connected to traceable evidence and known limitations	Grounded claim or explicitly bounded hypothesis	Fluent text or plausible structures may conceal unsupported relations	Citation verification, activity-cliff tests, local counterexamples, prospective experiments, and accountable human review

Limitations and Boundary Conditions

The proposed architecture is a conceptual construction rather than an empirically validated model. Its distinction between shared and modality-specific semantics requires operational definitions that may vary by task, biological scale, measurement process, and intended decision. Some relations may be shareable for retrieval but unsafe for generation or mechanistic interpretation. The architecture does not determine automatically which ontology, tokenizer, latent geometry, fusion mechanism, or confidence method is optimal, and it cannot guarantee that its conceptual components will be identifiable within trained representations.

The evidence base is heterogeneous in modality, task, dataset, and evaluation practice. Several source studies establish feasibility within a particular representational pair or benchmark rather than proving general interoperability across all modalities. Paired data may also be biased toward well-studied compounds, proteins, assays, instruments, cell systems, and publications. Scientific text introduces reporting and citation biases, while spectra and images retain laboratory-specific acquisition effects. Consequently, an implementation may inherit systematic gaps even when its internal alignment objectives are optimized successfully.

The architecture cannot establish molecular mechanism, causality, synthesizability, developability, safety, therapeutic benefit, clinical utility, regulatory acceptability, or operational readiness. A shared representation may support hypothesis generation or evidence organization, but it must not replace orthogonal measurements, perturbational validation, medicinal-chemistry assessment, or accountable scientific judgment. Misuse is especially likely when inferred modalities are presented as observations, latent similarity is described as explanation, or benchmark gains are generalized beyond the evaluated domain.

Research Agenda and Implementation Priorities

The first priority is to construct provenance-resolved multimodal evaluation resources in which identity, form, sample, perturbation, dose, time, acquisition process, and source are explicitly linked. These resources should contain both matched observations and scientifically meaningful hard negatives. Evaluation designs should vary modality availability, introduce controlled contradictions, preserve local chemical neighborhoods, and test whether private modality information survives alignment. Prospective studies should examine whether cross-modal representations improve defined scientific tasks without obscuring failure or encouraging unsupported transfer.

A second priority is development of reporting standards for multimodal molecular systems. Reports should state the semantic relation used for each alignment, distinguish observed from inferred modalities, describe negative sampling and data leakage controls, document uncertainty and abstention behavior, and separate native-modality tests from downstream task performance. Infrastructure should support versioned identifiers, source-linked textual assertions, machine-readable provenance, instrument and assay metadata, uncertainty ledgers, and reproducible transformations between raw records and model inputs.

Implementation should proceed in stages. Early systems should focus on bounded retrieval and evidence organization, where relationships can be inspected against source records. Later stages may evaluate controlled transfer, editing, or generation only after identity preservation, modality-specific fidelity, contradiction sensitivity, and uncertainty behavior are established.

Movement toward consequential pharmaceutical use should require external validation, prospective monitoring, expert review, and explicit stopping conditions. This staged agenda is intended to expose what remains unvalidated rather than to imply that a universal implementation is presently attainable.

Conclusion

One molecular language should not mean one undifferentiated encoding of every pharmaceutical data source. The proposed architecture instead defines a controlled representational commons in which chemical structures, biological sequences, three-dimensional conformations, spectra, images, and scientific text can share defensible semantic anchors while preserving their native information, provenance, uncertainty, and contradictions. Its contribution is the organization of identity, alignment, private residuals, missingness, uncertainty, and task-specific evidence gates into a coherent but explicitly unvalidated architecture. The decisive test is not whether modalities can be placed near one another in a latent space, but whether the resulting system preserves scientifically meaningful distinctions and fails transparently when evidence is absent, conflicting, or outside scope. Such an architecture may support more disciplined retrieval, generation, transfer, and grounding, but its value remains conditional on empirical validation, experimental confirmation, and accountable pharmaceutical interpretation.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. 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
2. 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
3. Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol. 2020;38(2):143-5. doi:10.1038/s41587-020-0418-2
4. Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, et al. Artificial intelligence foundation for therapeutic science. Nat Chem Biol. 2022;18(10):1033-6. doi:10.1038/s41589-022-01131-2
5. 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
6. Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nat Commun. 2022;13(1):2453. doi:10.1038/s41467-022-29939-5
7. Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, et al. Data-analysis strategies for image-based cell profiling. Nat Methods. 2017;14(9):849-63. doi:10.1038/nmeth.4397
8. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, et al. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. 2019;16(4):299-302. doi:10.1038/s41592-019-0344-8
9. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572-83. doi:10.1021/acscentsci.9b00576
10. 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
11. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15). doi:10.1073/pnas.2016239118
12. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9. doi:10.1038/s41586-021-03819-2
13. Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423-43. doi:10.1109/TPAMI.2018.2798607
14. Zeng Z, Yao Y, Liu Z, Sun M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. Nat Commun. 2022;13(1):862. doi:10.1038/s41467-022-28494-3
15. 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
16. Liu S, Nie W, Wang C, Lu J, Qiao Z, Liu L, et al. Multi-modal molecule structure–text model for text-based retrieval and editing. Nat Mach Intell. 2023;5(12):1447-57. doi:10.1038/s42256-023-00759-6

17. 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
18. Bushuiev R, Bushuiev A, Samusevich R, Brungs C, Sivic J, Pluskal T. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. Nat Biotechnol. 2026;44(4):630-40. doi:10.1038/s41587-025-02663-3
19. Seal S, Carreras-Puigvert J, Trapotsi MA, Yang H, Spjuth O, Bender A. Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. Commun Biol. 2022;5(1):858. doi:10.1038/s42003-022-03763-5
20. Haghighi M, Caicedo JC, Cimini BA, Carpenter AE, Singh S. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. Nat Methods. 2022;19(12):1550-7. doi:10.1038/s41592-022-01667-0
21. Xiong Z, Wang Z, Huang F, Qiu M, Fang S, Yang L, et al. Multi-to-uni modal knowledge transfer pre-training for molecular representation learning. Nat Commun. 2026;17(1):3797. doi:10.1038/s41467-026-69302-6
22. Arevalo J, Su E, Ewald JD, van Dijk R, Carpenter AE, Singh S. Evaluating batch correction methods for image-based cell profiling. Nat Commun. 2024;15(1):6516. doi:10.1038/s41467-024-50613-5
23. 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
24. Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770-80. doi:10.1021/acs.jcim.0c00502
25. 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
26. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096-108. doi:10.1021/acs.jcim.8b00839
27. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular Sets (MOSES): A benchmarking platform for molecular generation models. Front Pharmacol. 2020;11:565644. doi:10.3389/fphar.2020.565644
28. van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model. 2022;62(23):5938-51. doi:10.1021/acs.jcim.2c01073