Specialty Journal of Pharmacognosy, Phytochemistry, and Biotechnology
ISSN: 3062-441X
2026, Volume 6, Issue 1, Page No: 40-49

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Natural-Product Spectra Need Models That Know When They Are Outside Domain: An Uncertainty-Calibrated Architecture for MS/MS- and NMR-Assisted Structure Prioritization

Santiago Morales[1]*, Laura Rojas[2], Camila Vega[1], Diego Molina[3]

[1] Department of Spectral Modeling and Uncertainty Calibration, Faculty of Pharmacy, University of Buenos Aires, Buenos Aires, Argentina.
[2] Department of MS/MS and NMR Structure Prioritization, Faculty of Pharmacy, Pontifical Catholic University of Chile, Santiago, Chile.
[3] Department of Out-of-Domain Spectral Detection, Faculty of Pharmacy, National University of La Plata, La Plata, Argentina.

***E-mail** ✉ santiago.morales@agro.uba.ar

Received: 16 November 2025; **Revised:** 05 February 2026; **Accepted:** 08 February 2026

ABSTRACT

Natural-product structure prioritization increasingly uses machine-learning models that map tandem mass spectra, nuclear magnetic resonance data, or combinations of spectral modalities to candidate structures. Yet a high model score does not establish that a query lies within the chemistry, acquisition conditions, spectral information regime, or candidate space represented during model development. This distinction matters especially for natural products, where scaffold novelty, sparse reference coverage, heterogeneous acquisition, and incomplete spectral evidence can coincide. This article develops an original computational architecture for uncertainty-calibrated structure prioritization. The analysis separates chemical familiarity from predictive uncertainty, defines spectral-model domain as multidimensional rather than purely similarity-based, decomposes uncertainty across measurement, formula inference, spectral prediction, learned representation, candidate generation, ranking, and calibration, and proposes explicit decision states for ranking, qualified acceptance, abstention, and human review. MS/MS and NMR are treated as complementary but non-interchangeable evidence sources whose uncertainties should remain visible before fusion. The architecture further proposes that domain status, candidate-set completeness, and calibration be evaluated separately and that benchmark design deliberately challenge models with unfamiliar chemistry and shifted analytical conditions rather than relying only on random held-out tests. The framework is conceptual rather than prospectively validated; its thresholds, fusion rules, abstention criteria, and transferability across instruments, chemical classes, and laboratories require empirical testing. Its intended contribution is therefore not a new performance claim, but a reliability-centered design logic for deciding when spectral predictions should be trusted, qualified, expanded, or withheld.

Keywords: Natural products, Tandem mass spectrometry, Nuclear magnetic resonance, Out-of-distribution detection, Uncertainty calibration, Structure prioritization

How to Cite This Article: Morales S, Rojas L, Vega C, Molina D. Natural-Product Spectra Need Models That Know When They Are Outside Domain: An Uncertainty-Calibrated Architecture for MS/MS- and NMR-Assisted Structure Prioritization. Spec J Pharmacogn Phytochem Biotechnol. 2026;6(1):40-9. https://doi.org/10.51847/Tn3lzGlxg6

Introduction

Tandem mass spectrometry has become a powerful computational entry point for small-molecule structure prioritization because fragment and isotope information can be translated into molecular-formula and structural constraints rather than treated only as a library-search fingerprint. SIRIUS 4 exemplifies this shift by converting MS/MS evidence into ranked metabolite-structure information through formula and fingerprint inference [1]. The important distinction is epistemic: such workflows reduce a candidate space, but the resulting rank is not equivalent to direct structural proof. This distinction becomes more consequential when the molecule under

investigation is poorly represented in spectral libraries or lies outside the chemistry on which a ranking system was developed.

Computational annotation can nevertheless extend beyond exact library membership. COSMIC demonstrated that metabolites absent from reference spectral libraries can still be prioritized by combining experimental spectra with in-silico candidate evidence [2]. That result weakens a closed-world assumption in which "not in the library" means "not computationally addressable," but it does not remove dependence on candidate generation, chemical coverage, or the reliability of the learned score. For natural-product analysis, the unresolved question is therefore not simply whether a model can return a candidate, but whether it can recognize conditions under which its ranking should carry reduced evidential weight.

NMR-based learning provides a complementary route. Multitask learning from routine one-dimensional NMR spectra can recover useful structural information and support automated elucidation [3]. More recent multimodal systems show that different raw spectral sources can be combined to constrain candidate structures jointly [4]. These developments create a reliability problem broader than either modality: a model may produce a numerically confident answer while the query is chemically unfamiliar, experimentally shifted, poorly informative, or absent from its effective candidate universe.

The architecture developed here consequently treats spectral structure prioritization as an open-set decision problem. A query may be answerable from available candidates, may require expansion of the candidate universe, or may be insufficiently supported for a defensible rank. These possibilities should remain visible rather than being hidden behind a single confidence value. The framing also separates computational usefulness from analytical finality: a model can materially reorder structural hypotheses without establishing chemical identity. Reliability is therefore treated as a property of the decision process, not merely of the top-ranked candidate. The proposed architecture keeps domain status, predictive uncertainty, candidate availability, and analytical validation distinct so that the system can express not only what it favors, but also when that preference should be qualified or withheld.

Why prediction confidence is not the same as chemical familiarity

Out-of-distribution status is not a single intrinsic property of a molecule. Real-world molecular OOD analyses show that conclusions depend on how distribution shift is specified and which portion of molecular space is regarded as development support [5]. Coverage itself is uneven: benchmark and database composition can overrepresent some regions while leaving others sparsely sampled [6]. A natural-product query can therefore appear familiar under one representation or neighborhood definition yet remain poorly supported under another, particularly when the relevant scaffold or combination of structural features is uncommon.

This is why predictive uncertainty should not be used as a surrogate for chemical familiarity. Work at the edge of chemical space shows that model behavior, uncertainty, and molecular distance can diverge rather than vary monotonically together [7]. More generally, molecular-model reliability and uncertainty must be evaluated explicitly under distribution shift rather than inferred from in-distribution performance [5, 8, 9]. A low uncertainty estimate may therefore coexist with poor domain support, whereas a high uncertainty estimate may arise for chemically familiar material because its spectrum is noisy, conflicting, or intrinsically nondiscriminating.

The converse mistake is to reduce domain to molecular similarity alone. Neural uncertainty estimators differ in behavior across molecular prediction settings [10], while similarity-based applicability-domain metrics can provide useful evidence about whether a query resembles represented training chemistry [11]. Yet local similarity does not guarantee easy prediction: activity-cliff analyses demonstrate that near-neighbor molecules can remain difficult when small structural changes produce sharp response differences [12]. For spectral prioritization, this is a methodological warning rather than evidence for a directly equivalent class of "spectral cliffs."

The distinction produces four proposed interpretive states: familiar and stable predictions, familiar but uncertain predictions, unfamiliar but apparently stable predictions, and unfamiliar uncertain predictions. The third is particularly important because numerical confidence can conceal extrapolation. These categories are not validated cutoffs. Their purpose is to prevent a confidence estimate from silently answering a different question about whether the model has relevant chemical precedent. Chemical familiarity asks whether the model is operating near represented evidence; uncertainty asks how stable and calibrated the present inference appears.

Defining domain in natural-product spectral models

For spectral models, domain should be defined at more than the molecular-graph level. In MS/MS, the same compound can yield materially different fragmentation patterns across analytical platforms, making acquisition conditions part of the effective domain rather than incidental metadata [13]. Likewise, deep models that predict electrospray tandem spectra learn relationships from represented chemical structures and spectra, making chemical support inseparable from the training distribution used to establish those mappings [14]. A spectrum can therefore correspond to familiar chemistry while remaining analytically shifted, or originate under conventional acquisition conditions while representing chemically remote structure space.

Representation introduces a second layer. Graph-transformer spectrum-prediction models learn structure-to-fragmentation mappings whose reliability remains conditioned by their training chemistry [15]. Large-scale self-supervised learning from millions of tandem mass spectra can broaden reusable spectral representations, but scale does not logically remove gaps in corpus coverage [16]. The proposed architecture therefore treats representation support as a distinct domain question: does the query occupy a region where the learned molecular or spectral representation has credible empirical precedent?

A third layer is spectral informativeness. Cross-ionization similarity modeling has shown that low-information spectra and spectra dissimilar to training data can be detected explicitly, demonstrating that measurement quality and model familiarity can be assessed as related but separable reliability signals [17]. A sparse or atypical spectrum should therefore not acquire apparent certainty merely because the underlying molecular structure would have been familiar under another representation.

NMR adds its own contextual domain. Large experimental NMR resources reveal the breadth and heterogeneity of available empirical spectra [18], while solvent-aware prediction demonstrates that solvent can alter the mapping between molecular structure and observed NMR features [19]. Experimental context must therefore remain part of domain definition rather than being absorbed into an undifferentiated model score.

Candidate-space support requires separate treatment because even a well-performing ranker cannot recover a structure that was never supplied or generated for ranking. It asks whether an appropriate molecular formula, connectivity class, stereochemical alternative, or sufficiently relevant generated hypothesis can enter the decision set. This differs from representation support: a query may lie close to familiar embeddings while the correct candidate is absent, or be representationally remote while a correct candidate is nevertheless available.

Accordingly, this article proposes five interacting domain dimensions: chemical-space support, acquisition/context compatibility, spectral informativeness, representation support, and candidate-space support. They are not asserted as a validated universal taxonomy. Rather, the classification prevents one favorable signal from masking another form of extrapolation and treats disagreement among dimensions as diagnostically informative instead of something to average away.

Table 1 organizes these domain dimensions by the reliability question each addresses and the failure that becomes possible when it is ignored.

Table 1. Proposed multidimensional domain classification for natural-product spectral structure prioritization

Domain dimension	Reliability question	Relevant evidence	Principal failure if ignored
Chemical-space support	Is related chemistry represented?	Structural neighborhood; scaffold precedent	Unsupported chemical extrapolation
Acquisition/context compatibility	Were comparable conditions represented?	Platform; ionization; solvent; experiment type	Measurement shift mistaken for structural evidence
Spectral informativeness	Does the observation discriminate candidates?	Fragment richness; peak pattern; usable correlations	Sparse evidence converted into false precision
Representation support	Is the query supported in learned feature space?	Molecular or spectral embedding neighborhood	Latent proximity treated as universal familiarity
Candidate-space support	Can the true structure enter the ranked set?	Formula constraints; library or generator coverage	Forced ranking when the truth is absent

Note: This classification is proposed as a reliability framework. The dimensions may interact and should not be interpreted as statistically independent or as validated decision thresholds.

Sources of uncertainty

Uncertainty enters before a final structure score is produced. Forward prediction on previously untrained tandem spectra exposes generalization error that may remain hidden when evaluation is concentrated on familiar examples [20]. Even when the measured spectrum is technically adequate, the scoring rule itself matters: spectral entropy

and conventional dot-product similarity can yield different retrieval behavior [21]. A candidate rank therefore inherits uncertainty both from the predictive model and from the metric used to compare predicted, experimental, or reference spectra.

Upstream molecular-formula inference adds another layer. MIST-CF demonstrates that formula candidates can be inferred from tandem spectra using learned and fragment-based evidence [22], while BUDDY uses bottom-up MS/MS interrogation to constrain formula discovery [23]. Such approaches can substantially reduce search space, but formula uncertainty propagates downstream; an incorrect upstream restriction can exclude the correct structure before structural ranking begins. Formula confidence should therefore remain distinguishable from structure-level confidence.

Learned representations introduce further uncertainty because distances and similarities in latent space are properties of a trained objective rather than direct chemical measurements. MSBERT organizes tandem spectra into a learned chemically meaningful representation [24], and contrastive spectrum–structure pretraining can align spectral and molecular representations for identification tasks [25]. These embeddings provide useful evidence, but their numerical proximity should not be interpreted as a calibrated probability of molecular identity without separate validation.

NMR-driven prioritization has parallel failure routes. Automated NMR systems depend on forward prediction, spectral matching, candidate generation, and the correspondence between simulated and experimental data [26]. Earlier automated NMR frameworks likewise narrow a difficult inverse problem rather than abolishing structural degeneracy [27]. Hybrid machine-learning and quantum-mechanical approaches can rank structural alternatives probabilistically, but those probabilities remain conditional on the candidate set and modeled observables [28].

These uncertainties need not be independent. A poor spectrum can destabilize formula inference, distort a learned representation, and enlarge candidate ambiguity simultaneously. The proposed architecture therefore does not add uncertainty components as though they were separable numerical errors. Instead, measurement, representation, candidate-generation, ranking, and calibration uncertainty remain traceable so that a downstream decision can identify which evidential layer is unsupported or unresolved.

These uncertainty pathways and their consequences for domain-aware structure decisions are integrated in **Figure 1.**

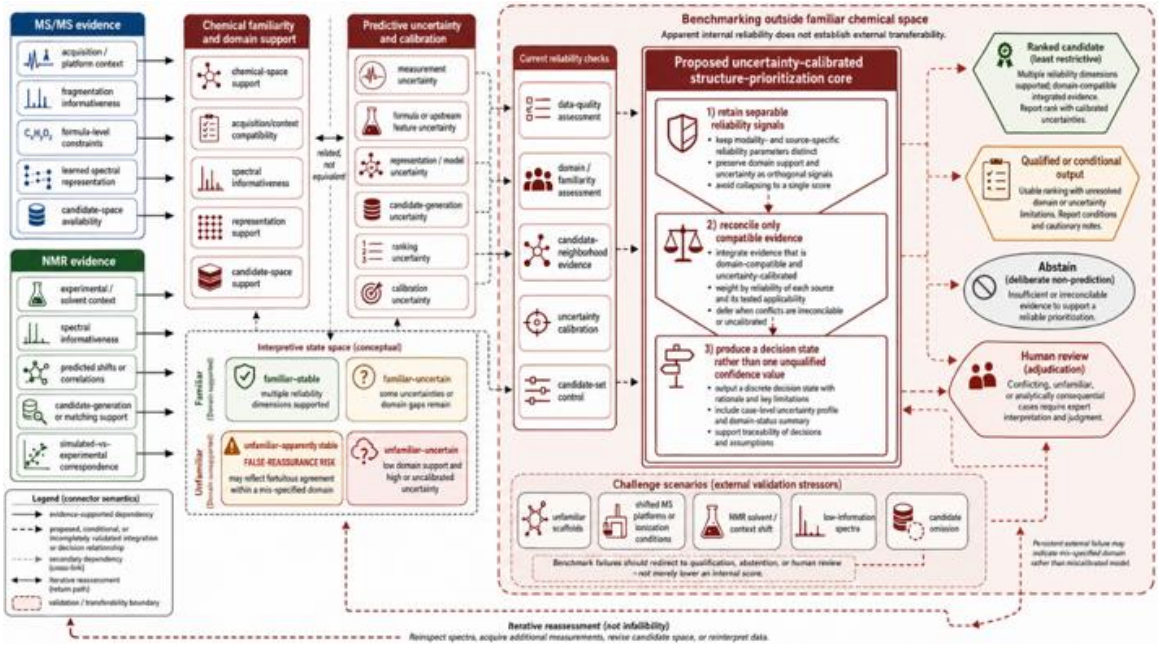

Figure 1. When a spectral model should rank, qualify, or defer: proposed domain-aware uncertainty architecture

Current strategies for detecting unreliable predictions

Current reliability strategies address different failure modes and should not be treated as interchangeable. Neighbourhood-aware retrieval is one example. MS2Query combines spectral similarity with additional learned

evidence to retrieve and rerank molecular analogues, showing that the local candidate neighbourhood can provide information beyond a single raw similarity score [29]. Such agreement is useful when relevant neighbours exist, but it cannot establish reliability when the reference neighbourhood is sparse or chemically inappropriate. Likewise, explicit spectral-quality assessment can flag spectra that are poorly informative or dissimilar to development data [17]; this is evidence about whether a query deserves caution, not proof that any particular candidate is wrong.

Other strategies expand rather than merely inspect the candidate space. Recent deep-learning approaches can infer molecular formula from tandem spectra [30], while structure-informed generation can propose candidates beyond a fixed library [31]. These methods reduce one form of closed-world dependence but introduce another reliability question: whether the newly admitted formula or generated structure is adequately supported by the measurement. Candidate expansion therefore should not be interpreted as confidence expansion.

The proposed taxonomy separates four reliability functions: domain assessment asks whether the query resembles supported chemistry and experimental context; data-quality assessment asks whether the measured spectrum carries sufficient discriminating information; uncertainty calibration asks whether predictive uncertainty corresponds to observed error behavior; and candidate-set control asks whether the decision space is plausibly complete enough for ranking. No current method establishes all four simultaneously. A robust workflow should preserve their outputs separately until a decision rule can explain why a candidate is ranked, qualified, or withheld.

A reliability signal is also more interpretable when its failure semantics are explicit. “Outside chemical support,” “poor spectrum,” “candidate set incomplete,” and “calibration unreliable” should not be collapsed into a generic low-confidence flag because they imply different remedial actions. The first may require caution about extrapolation; the second may motivate reacquisition; the third may require candidate expansion; and the fourth may require recalibration or withholding probabilistic interpretation.

Table 2 compares these strategies by the reliability question they can address and, equally importantly, the conclusion they cannot justify.

Table 2. Reliability strategies for spectral structure prioritization and their non-equivalent evidential roles

Strategy	Primary question	Useful signal	Does not establish
Domain assessment	Is the query supported by represented chemistry/context?	Similarity, coverage, acquisition compatibility	That the top candidate is correct
Spectral-quality assessment	Is the measurement sufficiently informative?	Fragment richness, atypicality, usable correlations	That familiar chemistry is in domain
Neighbourhood/reranking evidence	Do related references support the same hypothesis?	Analogue agreement, reranked similarity	Completeness of the candidate universe
Uncertainty calibration	Does stated uncertainty track observed error?	Calibrated intervals, sets, reliability statistics	Chemical familiarity by itself
Candidate-set control	Could the true structure plausibly enter consideration?	Formula constraints, library expansion, generation	Analytical validation of generated candidates

Note: The separation is proposed; individual methods may contribute to more than one function.

An uncertainty-calibrated structure-prioritization architecture

The proposed architecture begins by refusing to collapse reliability into one confidence score. Before candidate ranking, a domain-state gate records chemical-space support, acquisition/context compatibility, spectral informativeness, representation support, and candidate-space support. These states may disagree. A chemically familiar query acquired under shifted conditions, for example, should not inherit the same reliability status as a familiar query measured under represented conditions. Conversely, an unfamiliar scaffold should remain visibly unfamiliar even when a model returns a narrow predictive distribution [7].

A second layer retains an uncertainty ledger rather than immediately averaging uncertainty sources. Formula-level ambiguity, forward spectral prediction error, representation instability, candidate-generation uncertainty, ranking uncertainty, and calibration error can affect different parts of the inference chain. Distribution shift can alter uncertainty quality itself [9]. The architecture therefore proposes that candidate scores remain accompanied by their evidential provenance: which spectral observations supported them, which transformations or learned models intervened, and which uncertainty component generated a warning.

A third layer treats candidate omission as a first-class possibility. Library-absent compounds can sometimes be ranked computationally [2], and generative methods can broaden candidate search [31], but neither result licenses a forced choice from a deficient set. The architecture assigns an explicit open-set state when candidate-space support is inadequate.

The gate is intentionally noncompensatory at critical boundaries. Strong agreement on one layer should not automatically cancel evidence that the query is outside another essential layer. A highly similar MS/MS neighbour, for example, does not repair an acquisition mismatch by arithmetic averaging, and a calibrated ranking among supplied structures does not repair candidate omission. The architecture therefore proposes typed warnings that can constrain downstream actions independently of the rank score. It also preserves a distinction between "insufficient evidence" and "conflicting evidence": both may prevent commitment, but only the latter presupposes informative observations that disagree.

Finally, set-valued decisions provide a practical alternative to unconditional top-one ranking. Conformal molecular retrieval from mass spectra has demonstrated spectrum-specific prediction sets under in-distribution and shifted settings [32]. This supports the feasibility of calibrated candidate sets, while not implying that conformal coverage itself diagnoses every form of chemical OOD. The proposed architecture therefore permits four outputs: a ranked candidate, a qualified ranking with unresolved limitations, abstention, or escalation for human adjudication. These states require prospective validation and application-specific operating criteria.

Figure 2 translates this proposal into a decision architecture in which calibration and modality evidence determine when ranking should remain conditional or give way to abstention.

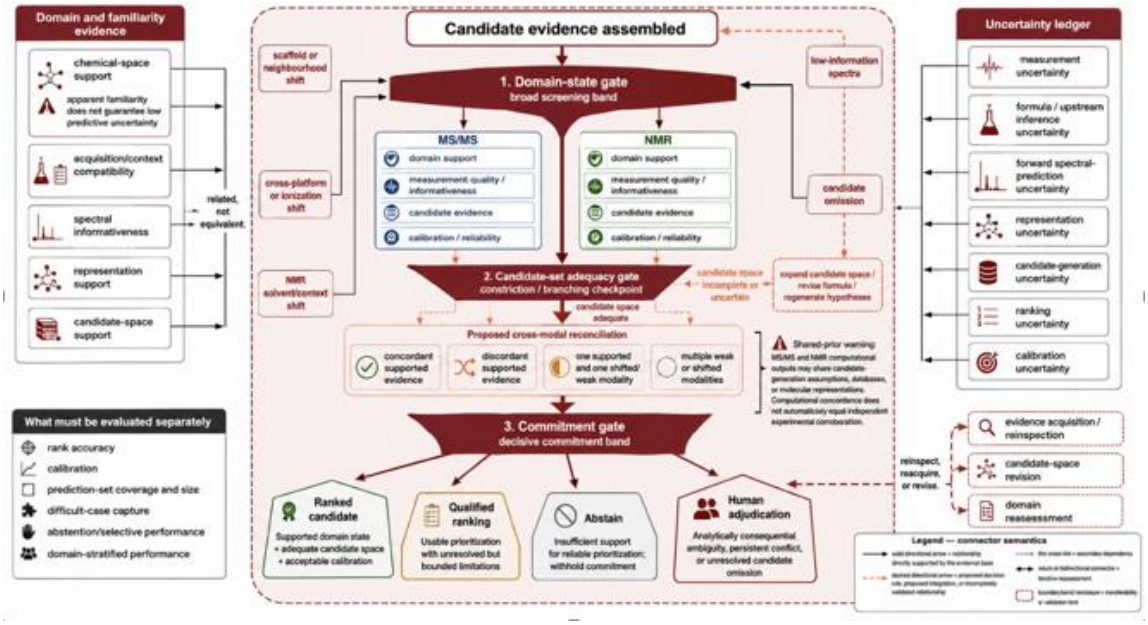

Figure 2. From candidate scores to defensible decisions: proposed calibration, fusion, and abstention gates

Combining evidence across spectral modalities

Cross-modal evidence is most useful when it reduces different ambiguities rather than merely increasing the number of model scores. Fast HSQC-based workflows can support natural-product dereplication by providing correlation patterns that constrain structural hypotheses [33]. Such NMR evidence is qualitatively different from MS/MS fragmentation evidence: agreement can strengthen prioritization, but disagreement should not be hidden by averaging two scores into a single apparently precise confidence value.

Calibration should therefore occur, where feasible, at the modality level before fusion. Uncertainty-calibrated NMR structure confirmation demonstrates that candidate assessment can be accompanied by an explicit estimate of reliability [34]. The principle proposed here is broader but unvalidated: MS/MS and NMR evidence should each retain its own domain status, quality assessment, candidate coverage, and calibration record before a fusion rule is applied. This makes it possible to distinguish "two supported modalities agree" from "one supported modality overwhelms one shifted or poorly informative modality."

Cross-modal dependence is another boundary. MS/MS and NMR may share upstream candidate-generation assumptions, database priors, or molecular representations, so concordant computational outputs are not necessarily independent confirmations. The proposed fusion layer therefore records whether agreement arises from genuinely orthogonal observations or from partially shared computational machinery. This prevents duplicated priors from being counted twice as if they were independent experimental support.

Flexible multispectral systems further show that structure elucidation need not assume an identical set of spectra for every query [35]. The architecture consequently treats missingness as a decision condition, not automatically as failure. A well-supported single modality may justify prioritization, whereas several weak or shifted modalities may justify abstention. Fusion should therefore be reliability-weighted in a qualitative architectural sense, without prespecifying numerical weights that have not been validated.

Table 3 expresses the proposed workflow as decision states rather than a universal fusion formula.

Table 3. Proposed decision logic for combining MS/MS and NMR evidence without erasing modality-specific uncertainty

Evidence state	Cross-modal relation	Candidate-space status	Proposed action
Both modalities supported	Concordant	Adequate	Rank; retain modality-specific evidence record
Both supported	Discordant	Adequate	Qualify; inspect conflict before commitment
One supported, one shifted/weak	Concordant or discordant	Adequate	Weight interpretation toward supported evidence; qualify
Evidence informative	Concordant	Incomplete	Expand candidate space; do not force top-one choice
Multiple weak/shifted sources	Any	Uncertain	Abstain or request additional measurement
Persistent unresolved conflict	Any	Adequate or uncertain	Escalate to human adjudication

Note: These are proposed decision states, not validated thresholds or mandatory operating rules.

Abstention and human intervention

Abstention is not the absence of a model output; it is an explicit decision that available evidence does not justify structural commitment at the requested level. Set-valued conformal retrieval provides one methodological route toward outputs that are broader than a forced top-one answer [32]. In the proposed architecture, however, abstention may also be triggered by inadequate domain support, low-information spectra, unresolved candidate omission, or disagreement between modalities. A spectral-quality warning, for example, identifies a reason to reconsider the evidential basis rather than simply lowering a candidate score [17].

The intervention rule is consequently multi-trigger rather than threshold-only. Chemical unfamiliarity can justify caution even when numerical uncertainty appears small [7], while calibrated NMR evidence may distinguish stronger from weaker candidate confirmation without making the calibration transferable to all chemistry [34]. The architecture proposes escalation when the unresolved issue is analytically consequential and additional interpretation could change the candidate universe or the evidence assigned to candidates.

Human intervention should not be represented as an infallible final classifier. Multimodal and automated NMR systems organize evidence and candidate hypotheses but remain bounded by the spectra, models, and search spaces supplied to them [4]. A human adjudication stage can instead audit raw data, question preprocessing, revise acquisition conditions, add orthogonal experiments, or introduce candidates that the computational system omitted. The workflow is therefore recursive: review may return the case to evidence acquisition, candidate generation, or domain assessment. Whether this improves reliability, efficiency, or reproducibility must be established prospectively rather than assumed.

Benchmarking outside familiar chemical space

A reliability architecture cannot be evaluated solely by random held-out accuracy. Realistic OOD definitions change the difficulty and meaning of molecular generalization tests [5], while explicit evaluation on untrained tandem spectra demonstrates why performance on familiar spectral distributions is insufficient [20]. Benchmarking should therefore stratify queries by the kinds of shift the architecture claims to recognize, instead of pooling all examples into one test statistic.

Chemical-space shift deserves explicit treatment. Models can behave differently near the edge of represented chemistry, OOD robustness can differ from ordinary predictive performance, and NMR model conclusions depend on standardized benchmark construction [7, 8, 36]. The benchmark should therefore preserve separate panels for represented chemistry, scaffold- or neighbourhood-shifted chemistry, and analytically shifted observations. For MS/MS, acquisition-platform differences are a documented source of spectral variation [13]; for NMR, solvent context can alter prediction behavior [19]. These challenges should be crossed where feasible because chemical and analytical shift may occur together in natural-product work.

Evaluation endpoints should also remain multidimensional. Calibration under distribution shift should be assessed independently of rank accuracy [9]. Spectral-quality or domain-warning systems should be judged by whether they capture difficult cases rather than by whether they merely correlate with candidate scores [17]. Set-valued approaches require coverage and set-size assessment in addition to conventional retrieval performance [32]. For an abstaining system, selective performance among retained predictions must be interpreted alongside how many and which cases are deferred.

Benchmark construction should moreover prevent easy cases from dominating summary metrics. Reporting performance separately across domain strata can reveal whether a model succeeds because most test molecules remain close to represented chemistry while failing in the very region where an OOD-aware architecture is intended to help. Challenge sets should retain enough metadata to attribute failure to chemistry, acquisition, spectral informativeness, candidate coverage, or combinations thereof. Without that attribution, an aggregate OOD label can obscure the mechanism of failure and provide little guidance for redesign.

The strongest test would be prospective external evaluation on unseen chemistry acquired under conditions not used for model development, with predefined procedures for candidate omission, multimodal disagreement, abstention, and expert adjudication. This proposal is deliberately stricter than ordinary random splitting. Failure under such testing could indicate miscalibration, inadequate domain specification, deficient candidate generation, or measurement shift; the benchmark should be designed to distinguish those explanations rather than reporting only an aggregate accuracy change.

Conclusion

Natural-product spectral models need a vocabulary for saying more than which candidate scores highest. The architecture proposed here separates chemical familiarity from predictive uncertainty and defines domain through interacting chemical, analytical, informational, representational, and candidate-space dimensions. It then preserves uncertainty across the inference chain, keeps MS/MS and NMR evidence distinguishable before fusion, and permits ranked, qualified, abstained, or human-adjudicated outcomes rather than forcing every spectrum into a single structural prediction.

This design is intended to change the object being calibrated. Reliability should attach to a traceable decision assembled from evidence, domain support, candidate availability, and uncertainty, not to an isolated top-score interpreted as structural certainty. An unfamiliar but numerically stable prediction therefore remains visibly extrapolative, while a chemically familiar but analytically poor spectrum can be withheld despite apparent molecular precedent.

The framework remains conceptual. Its domain variables may be correlated, its decision states require operational definitions, and no universal abstention threshold, fusion weight, or candidate-set criterion is established here. External testing must challenge unfamiliar chemistry and shifted measurement conditions together and determine whether the proposed decomposition improves failure detection without discarding useful predictions unnecessarily. Until such validation exists, the architecture should be read as an evidence-bounded design hypothesis: spectral models should be rewarded not only for ranking plausible structures, but for recognizing when the available evidence does not support confident structural commitment.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. 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
2. Hoffmann MA, Nothias LF, Ludwig M, Fleischauer M, Gentry EC, Witting M, et al. High-confidence structural annotation of metabolites absent from spectral libraries. Nat Biotechnol. 2022;40(3):411-21. doi:10.1038/s41587-021-01045-9
3. Hu F, Chen MS, Rotskoff GM, Kanan MW, Markland TE. Accurate and efficient structure elucidation from routine one-dimensional NMR spectra using multitask machine learning. ACS Cent Sci. 2024;10(11):2162-70. doi:10.1021/acscentsci.4c01132
4. Mirza A, Patiny L, Jablonka KM. End-to-end multimodal structure elucidation from raw spectra combining contrastive learning and evolutionary algorithms. Nat Commun. 2026;17(1):5013. doi:10.1038/s41467-026-73846-y
5. Tossou P, Wognum C, Craig M, Mary H, Noutahi E. Real-world molecular out-of-distribution: Specification and investigation. J Chem Inf Model. 2024;64(3):697-711. doi:10.1021/acs.jcim.3c01774
6. Kretschmer F, Seipp J, Ludwig M, Klau GW, Böcker S. Coverage bias in small molecule machine learning. Nat Commun. 2025;16(1):554. doi:10.1038/s41467-024-55462-w
7. van Tilborg D, Rossen L, Grisoni F. Molecular deep learning at the edge of chemical space. Nat Mach Intell. 2026;8(4):575-87. doi:10.1038/s42256-026-01216-w
8. Fooladi H, Vu TNL, Mathea M, Kirchmair J. Evaluating machine learning models for molecular property prediction: Performance and robustness on out-of-distribution data. J Chem Inf Model. 2025;65(19):9871-91. doi:10.1021/acs.jcim.5c00475
9. Parrondo-Pizarro R, Lanini J, Rodríguez-Pérez R. Uncertainty quantification in molecular machine learning for property predictions under data shifts. J Chem Inf Model. 2026;66(2):923-35. doi:10.1021/acs.jcim.5c02381
10. 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
11. Liu R, Wallqvist A. Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. J Chem Inf Model. 2019;59(1):181-9. doi:10.1021/acs.jcim.8b00597
12. 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
13. Hoang C, Uritboonthai W, Hoang LT, Billings EM, Aisporna A, Nia FA, et al. Tandem mass spectrometry across platforms. Anal Chem. 2024;96(14):5478-88. doi:10.1021/acs.analchem.3c05576
14. Chen B, Li H, Huang R, Tang Y, Li F. Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules. Nat Commun. 2024;15(1):8396. doi:10.1038/s41467-024-52805-5
15. Young A, Röst H, Wang B. Tandem mass spectrum prediction for small molecules using graph transformers. Nat Mach Intell. 2024;6(4):404-16. doi:10.1038/s42256-024-00816-8
16. 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
17. de Jonge NF, Chekmeneva E, Schmid R, Joas D, Truong LJ, van der Hooft JJJ, et al. Cross ionization mode chemical similarity prediction between tandem mass spectra in metabolomics. Nat Commun. 2026;17(1):2483. doi:10.1038/s41467-026-69083-y
18. Wang JJ, Jin Y, Zhi CY, Liu YJ, Huang XH, Xu F, et al. NMRexp: A database of 3.3 million experimental NMR spectra. Sci Data. 2025;12(1):1954. doi:10.1038/s41597-025-06245-5
19. Li Y, Xu H, Kumar A, Wang DS, Heiss C, Azadi P, et al. TransPeakNet for solvent-aware 2D NMR prediction via multi-task pre-training and unsupervised learning. Commun Chem. 2025;8(1):51. doi:10.1038/s42004-025-01455-9

20. Bremer PL, Vaniya A, Kind T, Wang S, Fiehn O. How well can we predict mass spectra from structures? Benchmarking competitive fragmentation modeling for metabolite identification on untrained tandem mass spectra. J Chem Inf Model. 2022;62(17):4049-56. doi:10.1021/acs.jcim.2c00936
21. Li Y, Kind T, Folz J, Vaniya A, Mehta SS, Fiehn O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. Nat Methods. 2021;18(12):1524-31. doi:10.1038/s41592-021-01331-z
22. Goldman S, Xin J, Provenzano J, Coley CW. MIST-CF: Chemical formula inference from tandem mass spectra. J Chem Inf Model. 2024;64(7):2421-31. doi:10.1021/acs.jcim.3c01082
23. Xing S, Shen S, Xu B, Li X, Huan T. BUDDY: Molecular formula discovery via bottom-up MS/MS interrogation. Nat Methods. 2023;20(6):881-90. doi:10.1038/s41592-023-01850-x
24. Zhang H, Yang Q, Xie T, Wang Y, Zhang Z, Lu H. MSBERT: Embedding tandem mass spectra into chemically rational space by mask learning and contrastive learning. Anal Chem. 2024;96(42):16599-608. doi:10.1021/acs.analchem.4c02426
25. Chen L, Xia B, Wang Y, Huang X, Gu Y, Wu W, et al. CMSSP: A contrastive mass spectra-structure pretraining model for metabolite identification. Anal Chem. 2024;96(42):16871-81. doi:10.1021/acs.analchem.4c03724
26. Jin Y, Wang JJ, Xu F, Ji X, Gao Z, Zhang L, et al. NMR-Solver: Automated structure elucidation via large-scale spectral matching and physics-guided fragment optimization. Nat Commun. 2026;17(1):4740. doi:10.1038/s41467-026-71315-0
27. Huang Z, Chen MS, Woroch CP, Markland TE, Kanan MW. A framework for automated structure elucidation from routine NMR spectra. Chem Sci. 2021;12(46):15329-38. doi:10.1039/D1SC04105C
28. Tsai YH, Amichetti M, Zanardi MM, Grimson R, Daranas AH, Sarotti AM. ML-J-DP4: An integrated quantum mechanics-machine learning approach for ultrafast NMR structural elucidation. Org Lett. 2022;24(41):7487-91. doi:10.1021/acs.orglett.2c01251
29. de Jonge NF, Louwen JJR, Chekmeneva E, Camuzeaux S, Vermeir FJ, Jansen RS, et al. MS2Query: Reliable and scalable MS2 mass spectra-based analogue search. Nat Commun. 2023;14(1):1752. doi:10.1038/s41467-023-37446-4
30. Hong Y, Li S, Ye Y, Tang H. FIDDLE: A deep learning method for chemical formulas prediction from tandem mass spectra. Nat Commun. 2025;16(1):11102. doi:10.1038/s41467-025-66060-9
31. Wang H, Zhang H, Zhu ZJ. Structure-informed deep generation enables de novo metabolite annotation in untargeted metabolomics. Nat Commun. 2026;17(1):5426. doi:10.1038/s41467-026-72149-6
32. Rakhshaninejad M, De Waele G, Jürgens M, Waegeman W. Reliable molecular retrieval from mass spectra using conformal prediction. J Chem Inf Model. 2026;66(10):5788-800. doi:10.1021/acs.jcim.6c00727
33. Yuan B, Zhang C, Ji C, Liu G, Li X, Gong S, et al. HSQCid: A powerful tool for paving the way to high-throughput structural dereplication of natural products based on fast NMR experiments. Anal Chem. 2025;97(6):3227-35. doi:10.1021/acs.analchem.4c03102
34. Kotlyarov R, Howarth A, Goodman JM. DP5 without DFT: Uncertainty-calibrated graph neural net accelerates structure confirmation via NMR. Chem Sci. 2026;17(18):9138-47. doi:10.1039/D5SC06988B
35. Priessner M, Lewis RJ, Lemurell I, Johansson MJ, Goodman JM, Janet JP, et al. Advancing structure elucidation with a flexible multi-spectral AI model. Angew Chem Int Ed. 2026;65(2). doi:10.1002/anie.202517611
36. Xu F, Guo W, Wang F, Yao L, Wang H, Tang F, et al. Toward a unified benchmark and framework for deep learning-based prediction of nuclear magnetic resonance chemical shifts. Nat Comput Sci. 2025;5(4):292-300. doi:10.1038/s43588-025-00783-z