



Machine Learning in Natural Product Separation and Purification Across Chromatographic Method Development, Peak Resolution, Fraction Prioritization, and Process Optimization

Maria Gonzales^{1*}, Jose Santos¹, Anna Cruz², Carlos Reyes³

¹Department of ML for Chromatographic Separation, Faculty of Pharmacy, University of the Philippines Manila, Manila, Philippines.

²Department of Peak Resolution and Fraction Prioritization, Faculty of Pharmacy, University of Santo Tomas, Manila, Philippines.

³Department of Process Optimization for Natural Product Purification, Faculty of Pharmacy, Ateneo de Manila University, Quezon City, Philippines.

ABSTRACT

Natural-product discovery fundamentally depends on the ability to separate chemically heterogeneous mixtures before assigning structures, activities, or mechanisms to individual constituents. Machine learning is increasingly applied to chromatographic method development, retention prediction, signal processing, fraction prioritization, and experimental optimization, yet these applications are often discussed as if they represented a single analytical problem. This methodological review distinguishes four different computational roles across the separation-to-purification continuum: predicting chromatographic behavior, interpreting chromatographic signals, prioritizing chemically informative or bioactive fractions, and controlling experimental search. The evidence indicates that retention and selectivity models are becoming more chemically and method aware, while peak-processing models can automate difficult signal-recognition tasks, and adaptive optimization can reduce unguided exploration of chromatographic conditions. Natural-product-specific studies further show that metabolomic features can support fraction prioritization before complete structural characterization. However, analytical prediction, signal resolution, fraction ranking, and preparative purification are not interchangeable validation levels. Generalization remains constrained by chemical-space shift, incomplete chromatographic metadata, instrument and column variation, preprocessing choices, objective specification, experimental budget, and sparse direct evidence connecting analytical machine-learning performance with preparative natural-product yield and purity. We therefore propose a layered methodological framework in which machine learning is evaluated according to the chromatographic object it learns, the decision it informs, and the experimental boundary across which it is expected to transfer, thereby supporting more rigorous comparison of methods while preserving the role of expert chromatography in defining objectives, diagnosing failure, and validating purification outcomes.

Key Words: Natural products, Machine learning, Chromatography, Purification, Retention prediction, Method development

eIJPPR 2025; 15(5):91-100

HOW TO CITE THIS ARTICLE: Gonzales M, Santos J, Cruz A, Reyes C. Machine Learning in Natural Product Separation and Purification Across Chromatographic Method Development, Peak Resolution, Fraction Prioritization, and Process Optimization. Int J Pharm Phytopharmacol Res. 2025;15(5):91-100. <https://doi.org/10.51847/aRslitHkQr>

INTRODUCTION

Natural products remain productive sources of pharmacologically relevant chemistry, but their discovery

value is inseparable from an analytical difficulty: extracts, fractions, and biosynthetic mixtures contain compounds whose abundance, polarity, ionization behavior, stereochemistry, and matrix context can differ markedly.

Corresponding author: Maria Gonzales

Address: Department of ML for Chromatographic Separation, Faculty of Pharmacy, University of the Philippines Manila, Manila, Philippines.

E-mail: ✉maria.gonzales@upm.edu.ph

Received: 29 May 2025; **Revised:** 28 September 2025; **Accepted:** 29 September 2025

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.



Contemporary natural-product discovery therefore depends not only on identifying interesting molecules but on producing sufficiently resolved chemical evidence to decide which signals, fractions, or compounds deserve interpretation [1]. Separation is not a neutral prelude to biology. It determines whether a measured activity can plausibly be localized, whether spectral features can be assigned without excessive ambiguity, and whether downstream structural work is directed toward a chemically coherent target.

Bioactivity-based molecular networking illustrates an intermediate analytical state between unresolved mixture and fully characterized active compound. Fraction-associated activity can be connected with molecular-network information before every constituent is structurally characterized, allowing chemical features to be prioritized during bioassay-guided fractionation [2]. Functional metabolomics extends this logic by linking metabolomic variation with phenotypic information, while retaining uncertainty about the molecular basis of the observed phenotype [3]. These approaches make separation and prioritization mutually informative rather than strictly sequential, but they do not erase the distinction between a correlated feature, a resolved constituent, a bioactive fraction, and a causally responsible molecule.

Machine learning enters this landscape at several different levels. It may predict retention or selectivity, recognize peaks or overlapping signals, rank fractions from incomplete chemical fingerprints, recommend new chromatographic conditions, or emulate process behavior. These uses should not be collapsed into a generic category of “AI-assisted purification.” This review instead treats chromatographic machine learning as a set of method-specific inference and control problems. Its central question is not merely whether an algorithm can reproduce a chromatographic output, but what analytical object has been learned, what chromatographic decision the output can legitimately support, and what additional experimental evidence is required before an analytical prediction is allowed to influence preparative purification.

Why separation problems are attractive to machine learning

Chromatographic separation is attractive to machine learning because method development is governed by interacting variables whose effects are difficult to isolate one at a time. Molecular structure, stationary-phase chemistry, mobile-phase composition, gradient shape, temperature, flow, loading, detector response, preprocessing, and the chosen definition of satisfactory separation can alter the optimization landscape. Contemporary liquid-chromatography research therefore treats artificial intelligence not as a single model class but as a family of approaches spanning prediction, experiment selection, and method optimization [4].

The methodological opportunity is strongest when the learning target is explicit. Retention prediction asks for a conditional estimate of chromatographic behavior. Peak models operate on measured signals and operational peak boundaries. Fraction-prioritization models rank incomplete chemical evidence against a biological or analytical objective. Closed-loop optimization uses observed experimental outcomes to select subsequent chromatographic conditions. These tasks are related, but successful prediction at one level does not validate the next. Accurate retention prediction does not demonstrate preparative resolution; accurate peak detection does not prove chemical purity; a bioactivity classifier does not identify a causal constituent; and an efficient optimizer cannot rescue a poorly chosen experimental objective.

A useful distinction is therefore between what is represented, what chromatographic state is imposed, what signal is observed, and what decision is ultimately requested from the model. Experimental verification connects these layers but prevents them from being treated as interchangeable validation endpoints. This review proposes that chromatographic machine learning should be interpreted through these coupled layers rather than through model architecture alone. The distinction also determines which evidence can legitimately be transferred from general analytical chromatography to natural-product purification.

The eligibility logic used to preserve the distinction between direct natural-product evidence and transferable chromatographic methodology is summarized in **Table 1**.

Table 1. Eligibility logic for evidence connecting machine learning with chromatographic separation and purification

Dimension	Included evidence	Qualified inclusion	Not treated as evidence
Publication	Peer-reviewed 2017–2025 journal article; DOI; applicable Q1 status	—	Preprints, conference-only reports, vendor material, duplicate versions
Task	Retention/selectivity; condition selection; peak processing; fraction prioritization; closed-loop or process optimization	Adjacent chromatographic modality with explicit transfer rationale	Generic AI without a chromatographic decision
Natural-product fit	Natural-product fractions, metabolomics, or purification	Small-molecule studies supporting transferable method principles	Non-natural-product results relabelled as natural-product validation

Validation	Defined retrospective, benchmark, experimental, prospective, or process evaluation	Limited validation with visible scope qualification	Prediction accuracy treated as purification success
Transfer	Chemical, method, instrument, matrix, or process context reported	Cross-domain evidence with explicit boundary	Unqualified transfer across modalities or scales
Claim	Endpoint matched to what was measured	Proposed synthesis identified as proposed	Unsupported causal, mechanistic, purity, yield, or readiness claims

Note: This manuscript-derived eligibility logic organizes a targeted methodological review; it is not a systematic-review flow diagram or quantitative evidence hierarchy.

The resulting methodological distinction is formalized in **Figure 1** as a proposed framework for locating where a machine-learning output sits between chemical description and chromatographic decision.

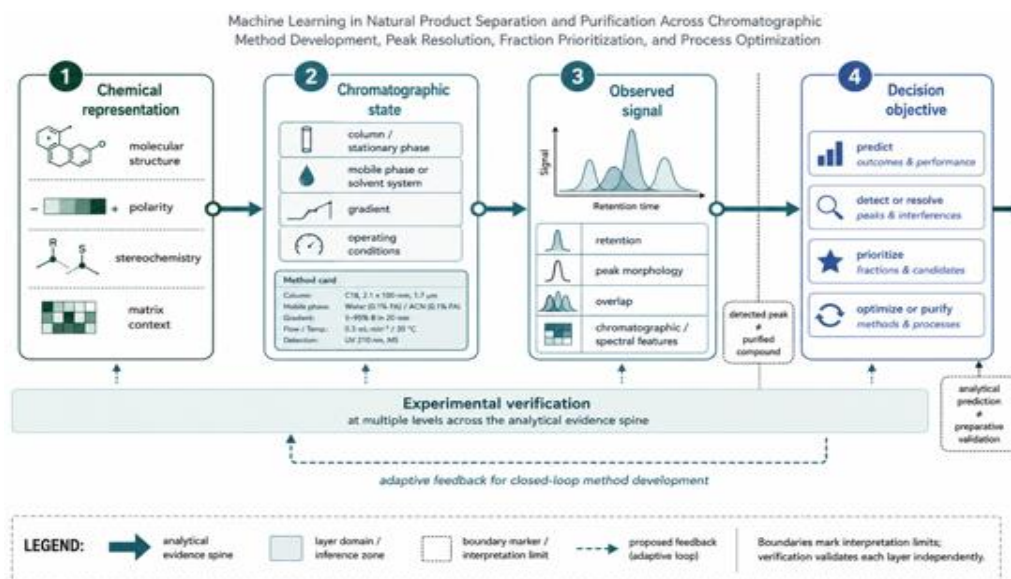


Figure 1. Four coupled inference layers in chromatographic machine learning

Review scope and methodological questions

The review addresses three linked methodological questions. First, what aspects of chromatographic separation are presently machine-learnable: molecular retention, selectivity, peak morphology, overlap, chromatographic fingerprints, fraction-level activity patterns, experimental response surfaces, or process behavior? Second, what type of decision does each model support—prediction, detection, ranking, condition selection, or process control? Third, how far can that decision travel beyond the chemical and chromatographic conditions under which the model was trained or evaluated?

Evidence is interpreted according to the distance between the demonstrated task and natural-product purification. Direct natural-product studies receive the greatest evidential weight for claims concerning fraction prioritization, metabolomic fingerprints, and bioactivity-guided separation. General small-molecule liquid-chromatography studies are retained where they establish a directly relevant methodological principle, but their outputs are not re-labelled as natural-product validation. Evidence from an adjacent chromatographic modality is

used only where the transferable methodological point is explicit and the modality boundary remains visible.

The review further separates analytical outputs from purification outcomes. Retention error, peak-detection performance, or optimization efficiency can establish performance for a defined analytical task; they do not independently establish recovered mass, fraction purity, yield, solvent efficiency, robustness under altered sample loading, or reproducibility at preparative scale. Accordingly, the review compares methods not only by predictive capability but by the decision they can support and the experimental evidence required to support the next stage of separation.

Search strategy and source selection

A targeted methodological search was used rather than an exhaustive systematic-review design. Eligible evidence was restricted to peer-reviewed journal articles published from 2017 through 2025 with a verifiable DOI and Q1 status in a relevant journal category for the applicable period. Searches covered natural-product separation and purification, retention and selectivity prediction, solvent and gradient selection, chromatographic peak detection

and deconvolution, LC–MS feature processing, fraction prioritization, active learning, Bayesian optimization, reinforcement learning, and chromatographic process modeling.

Selection required an exact role in a manuscript claim. Studies were excluded when they were adjacent to chromatography but could not support a defined separation, signal-processing, prioritization, optimization, validation, or transferability proposition. Duplicate versions, non-journal material, sources outside the eligible publication window, and evidence whose use would require unqualified extrapolation beyond its tested chemical, analytical, or process domain were not retained. Where direct natural-product evidence was unavailable, transferable chromatographic evidence was incorporated with an explicit boundary qualification rather than treated as equivalent validation.

Study characterization and method coding

Included studies were coded by the analytical object learned, input representation, chromatographic modality, model role, experimental design, comparator, validation scheme, reported outcome, uncertainty source, and transferability boundary. This coding separates models that learn molecular behavior from models that learn signal morphology or experimental response. It also distinguishes retrospective prediction from prospective experiment selection and process-level optimization.

Four methodological evidence roles emerged from this coding. Property-prediction studies estimate retention or selectivity from molecular and method information. Signal-inference studies detect, curate, quantify, or deconvolve chromatographic features. Prioritization studies use incomplete chemical fingerprints or metabolomic patterns to rank fractions or candidate chemistry. Control studies select chromatographic experiments or emulate process behavior. Natural-product relevance was coded independently of computational sophistication, because a technically strong model trained on drug-like molecules, environmental contaminants, or analytical test mixtures cannot automatically validate performance in chemically heterogeneous natural-product matrices. The subsequent synthesis therefore compares approaches according to the chromatographic decision they can support and the boundary conditions attached to that decision rather than according to headline model performance alone.

Predicting retention and selectivity

Retention prediction is the most mature machine-learning problem in this evidence set. Recent deep-learning work shows that retention time can be learned from molecular representations, but reported performance remains

inseparable from dataset composition, chromatographic conditions, and validation design [5]. Large experimental resources such as the METLIN retention-time dataset made supervised learning possible across much broader chemical spaces, while simultaneously exposing heterogeneity in the conditions under which retention measurements are generated [6]. RepoRT further demonstrates why reusable chromatographic prediction requires method metadata rather than retention values alone [7].

Graph neural networks offer a chemically natural representation because molecular connectivity can be learned without reducing structures to fixed handcrafted descriptors [8]. Interpretability studies indicate that graph-convolutional models can additionally expose learned structure–retention associations, although those associations should not be presented as mechanistic explanations of chromatography [9]. Selectivity creates an even stronger representation requirement: geometry-enhanced graph learning for enantioseparation shows that stereochemical information may alter prediction where constitutionally similar molecules behave differently [10]. Data scarcity and transfer remain central limitations. Pretraining can improve performance when the target retention dataset is small [11], while evaluation in industrial campaigns shows that apparently robust models remain conditioned by the chemical spaces on which they are deployed [12]. Cross-mode graph-transformer work suggests that information can sometimes transfer between reversed-phase chromatography and HILIC [13]. Condition-aware models likewise show that mobile-phase composition and stationary-phase context can be incorporated explicitly rather than treating retention as an intrinsic molecular property [14]. Collectively, these studies support a methodological shift from “predict the molecule's retention” toward predict chromatographic behavior conditional on molecule, method, and deployment domain.

Solvent-system and gradient optimization

Prediction becomes more directly actionable when the output concerns experimental conditions. Automated TLC combined with machine learning has been used to map chemical structure to polarity-related chromatographic behavior, providing a data-driven basis for narrowing solvent choices [15]. More directly, machine learning has been applied to column-chromatography condition prediction in defined chemical spaces [16]. Statistical and machine-learning analysis has also demonstrated explicit relationships between TLC observations and subsequent column conditions [17].

These studies provide an important bridge between molecular prediction and purification planning, but they do

not justify a universal solvent-selection rule. Matrix composition, loading, stationary-phase characteristics, compound abundance, and target purity can change the practical value of a predicted condition. For natural products, the defensible role of such models is therefore condition prioritization followed by chromatographic verification, not replacement of experimental method development.

Peak resolution as a multivariable prediction problem

Machine learning also operates downstream of physical separation, where the problem is no longer which condition to choose but what signal the chromatogram contains. Automated tuning of denoising algorithms illustrates that preprocessing can materially shape the signal supplied to subsequent models [18]. Convolutional neural networks have been applied directly to peak detection in reversed-phase LC [19], and PeakBot demonstrates that chromatographic peak picking can be formulated as a learned recognition problem rather than a collection of manually adjusted thresholds [20].

This changes the interpretation of “resolution.” A model may improve recognition of weak, asymmetric, or partially obscured features without improving the physical separation itself. The manuscript therefore distinguishes signal resolution from chemical resolution: the former concerns computational interpretation of measured data; the latter concerns whether distinct constituents have actually been separated sufficiently for the intended analytical or preparative decision.

Detecting co-elution and hidden chemical complexity

High-resolution LC–MS provides a particularly demanding setting because chemically meaningful features may be weak, crowded, or embedded in complex backgrounds. Deep learning has been used for precise peak detection in high-resolution LC–MS data [21], and transformer-based modeling now extends this approach toward combined peak detection and quantification in metabolomic workflows [22]. These methods increase the capacity to recognize structured signals that conventional processing may miss, but feature recognition is still not equivalent to proof that each signal corresponds to a uniquely resolved compound.

Explicit learned deconvolution of overlapping chromatographic–mass-spectral peaks has also been demonstrated with GC–MS data [23]. That result supports the broader methodological proposition that overlap can be treated as a learnable signal-unmixing problem, but it is not direct validation for LC-based natural-product co-elution. Deep-learning-assisted curation of large LC–MS datasets provides a further reminder that apparent chemical complexity depends partly on how candidate peaks are

accepted or rejected computationally [24]. Hidden complexity is therefore simultaneously a separation problem, a detection problem, and a data-curation problem.

Fraction prioritization before full structural characterization

Natural-product workflows rarely require complete structural elucidation of every detected constituent before deciding what to pursue. Bioactivity-based molecular networking already demonstrates that fraction-level biological information can be aligned with chemical-network evidence during iterative fractionation [2]. More recently, machine-learning classification using LC–MS/MS metabolomics has directly supported bioactivity prioritization of natural-product samples [25].

The methodological advantage is triage: chromatographic and spectral patterns can identify fractions or chemical neighborhoods that merit purification effort before every structure is known. The limitation is equally important. A model that assigns a high bioactivity probability has not established which constituent is responsible, whether multiple constituents interact, whether an assay artifact contributes to the classification, or what molecular mechanism is involved. Fraction prioritization should therefore reduce the search space for purification, not collapse prioritization into structural or pharmacological proof.

Chromatographic fingerprints as machine-learnable objects

Feature-based molecular networking provides a representation in which chromatographic and tandem-MS features can be retained as analyte-level objects even when structural annotation is incomplete [26]. Natural-product-oriented workflows such as NP3 similarly organize untargeted LC–MS evidence into reproducible feature-processing pipelines suitable for downstream comparison and prioritization [27].

This supports a broader methodological interpretation of the chromatographic fingerprint. A useful machine-learning object need not consist exclusively of fully identified molecules; it may instead encode retention, abundance, feature co-occurrence, spectral similarity, peak morphology, or fraction distribution. Such representations preserve information that would be lost by restricting analysis to confidently annotated compounds, but their validity depends on alignment, preprocessing, metadata, and platform reproducibility.

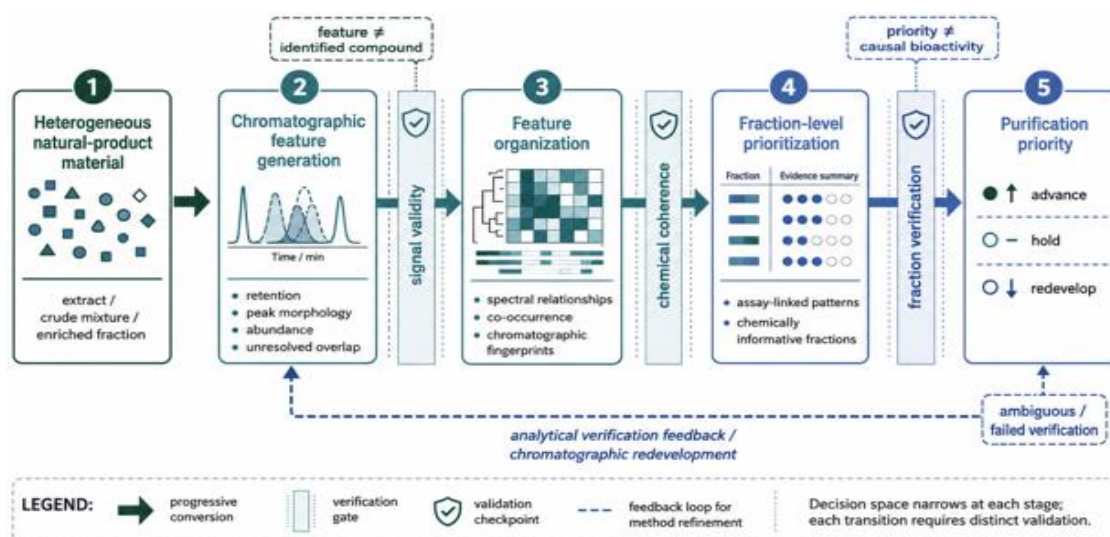
Table 2 summarizes how the evidence reviewed to this point changes meaning as the learned object moves from molecular structure to chromatographic signal and fraction-level fingerprints.

Table 2. Methodological domains represented in machine-learning studies of chromatographic separation

Learned object	Typical evidence input	Supported decision	Natural-product relevance	Principal boundary
Molecular behavior	Structure, geometry, method conditions	Retention/selectivity prediction	Transferable methodology	Domain and method shift
Separation conditions	TLC/column observations, solvent context	Condition prioritization	Emerging direct relevance	Matrix, loading, scale
Chromatographic signal	Peak shape, intensity, baseline, MS data	Detection/quantification	Direct analytical relevance	Detected feature ≠ pure compound
Overlapping signal	Composite chromatographic–spectral response	Deconvolution	Methodologically relevant	Modality-specific validation
Fraction fingerprint	LC–MS/MS features plus assay response	Fraction prioritization	Direct NP relevance	Activity class ≠ causal constituent
Feature network	Retention/spectral relationships	Dereplication and chemical triage	Direct NP relevance	Similarity ≠ structural identity

Abbreviation: NP, natural product.

The relationship among these representations is synthesized in **Figure 2**.



Closed-loop method development

Closed-loop chromatography changes the role of machine learning from prediction to experiment selection. Assisted active learning has been used in model-based LC method development to direct experimental effort toward informative conditions rather than uniformly exploring the available design space [28]. Bayesian optimization extends this principle by using measured chromatographic outcomes to update the model and select subsequent gradients [29]. Computer-driven optimization has also been demonstrated for more complex LC×LC gradient problems [30].

The central methodological change is feedback. Once experimental results alter the next condition selected, model validity depends not only on prediction error but on the objective function and acquisition strategy. Multi-task

Bayesian optimization illustrates how multiple separation objectives can be considered jointly [31]. Comparisons among optimization algorithms further show that apparent efficiency depends on the chromatographic problem and available experimental budget [32]. Reinforcement-learning studies make the same issue explicit: reward specification and experimental budget alter the behavior of automated method development [33]. Consequently, closed-loop automation does not remove expert judgment; it moves part of that judgment into the choice of objective, search space, constraints, reward, and stopping rule.

Scale-up from analytical separation to preparative purification

The weakest evidence transition is from analytical optimization to preparative natural-product purification.

Physics-based deep neural networks have been used to simulate and optimize nonlinear chromatographic processes [34], demonstrating that machine learning can represent process behavior rather than only analytical retention. Yet this does not establish preparative natural-product performance.

Column-condition prediction [16] and TLC-to-column translation [17] provide closer experimental bridges, but scale introduces additional variables: sample loading, column dimensions, mass transfer, matrix interference, solvent consumption, recovery, fraction collection, and purity requirements. The proposed methodological implication is therefore an analytical-to-preparative validation ladder. Predictions should be revalidated when the decision moves from retention, to analytical resolution, to fraction collection, and finally to preparative recovery. No preceding model-performance metric substitutes for the experimental endpoint at the new scale.

Process optimization under yield–purity trade-offs

Preparative purification also changes the definition of success. Analytical optimization often rewards resolution or peak separation, whereas process decisions may require simultaneous consideration of purity, recovery, yield, throughput, and solvent burden. Multiobjective chromatographic optimization demonstrates that different objectives can be modeled jointly [31], but the weighting of those objectives remains a scientific and operational decision.

Experimental budget also constrains attainable optimization [32]. For natural-product purification, an algorithm that produces a marginal resolution gain at excessive solvent or sample cost may be less useful than a robust condition with slightly lower analytical separation. Process-level models such as physics-informed chromatographic networks [34] make such trade-offs computationally approachable, but direct prospective evidence connecting these approaches with natural-product yield–purity performance remains limited. Any future process model should therefore be judged against measured purification outcomes rather than analytical surrogates alone.

Generalization across columns, instruments, and laboratories

Generalization is not a single property of a chromatography model. It can fail because the molecules, stationary phase, mobile phase, chromatographic mode, instrument, detector, preprocessing workflow, or laboratory differs from training conditions. Evidence from reusable retention repositories, industrial prediction campaigns, and cross-mode graph-transformer studies indicates that transfer depends on the relationship between the development domain and the intended chromatographic application rather than on molecular representation alone [7, 12, 13].

Condition-aware retention modeling reinforces this point because stationary- and mobile-phase variables contribute directly to prediction [14]. A model that appears chemically transferable may still be method-specific. For natural products, an additional matrix layer is present: complex extracts can contain unknown compounds and concentration regimes not represented in training sets. External validation should therefore report not merely whether accuracy declined, but which domain changed and whether the resulting error alters the chromatographic decision.

Data quality, reproducibility, and model interpretability

Large datasets improve learning opportunities but do not compensate for poorly described chromatography. Machine-learnable evidence requires sufficiently structured retention records, chromatographic metadata, and reproducible feature-processing workflows; dataset size alone does not remove those requirements [6, 7, 27]. Missing information on stationary phase, gradient, temperature, preprocessing, or detector conditions can convert methodological heterogeneity into apparent model noise.

Interpretability presents a related risk. Graph-convolutional models can identify molecular features associated with predicted retention [9], but those associations should not be reported as causal physicochemical mechanisms without independent evidence. At the signal level, learned denoising and curation introduce additional model-dependent transformations before prediction or prioritization occurs [24]. Reproducibility therefore concerns the whole analytical pipeline, not only the final model.

Table 3 integrates the major methodological failure modes with the validation response required before transfer.

Table 3. Evidence-synthesis matrix for validation failures in chromatographic machine learning

Failure source	What may change	Consequence	Required validation response	Evidence anchor
Chemical-space shift	Structures/stereochemistry	Retention or selectivity error	External chemical-domain test	[12]
Method shift	Phase, solvent, mode	Systematic retention change	Condition-aware evaluation	[14]
Signal shift	Noise, baseline, peak shape	Missed/spurious peaks	Independent signal benchmark	[18]

Processing shift	Feature extraction/curation	Altered fingerprint	Pipeline-level reproducibility test	[24]
Instrument/laboratory shift	Hardware and operating context	Calibration/domain drift	Cross-site validation	[7]
Scale shift	Loading/process conditions	Analytical success fails preparatively	Preparative purity/recovery validation	[34]

Where automation can—and cannot—replace expert chromatography

Automation is strongest where the chromatographic objective can be measured repeatedly and translated into an explicit optimization target. Closed-loop evidence shows that the optimizer, objective or reward, and available experimental budget jointly determine automated search behavior [29, 32, 33]. Human expertise therefore remains upstream of much apparently autonomous chromatography.

Experts still define chemically meaningful search spaces, decide whether a distorted peak represents overload or an analyte, recognize unexpected decomposition or irreversible adsorption, and determine whether a numerically improved objective is useful for purification. Active learning can reduce unnecessary experimentation [28], but expert oversight remains especially important in natural-product matrices containing unknown chemistry outside the model's training domain. The appropriate boundary is not “human versus machine chromatography.” It is which decisions can be formalized reliably, which require experimental verification, and which remain contingent on contextual chemical judgment.

Review limitations

This review is targeted and methodological rather than exhaustive. The evidence base is uneven: retention prediction, peak processing, and analytical method optimization are considerably better represented than direct machine-learning studies of preparative natural-product purification. Some transferable evidence therefore originates from pharmaceutical small molecules, contaminants, synthetic organic compounds, GC–MS, or process chromatography rather than complex natural-product extracts.

Performance metrics also differ enough across datasets and tasks that a numerical ranking of algorithms would imply comparability not established by the evidence. In particular, GC–MS overlap resolution [23] and nonlinear process modeling [34] are used to support methodological possibilities, not direct natural-product LC validation. Finally, the proposed distinctions between representation, signal inference, prioritization, and process control organize the reviewed evidence but have not themselves been prospectively validated as a universal framework.

The resulting research requirements are summarized in

Table 4.

Table 4. Proposed quality and research agenda for machine-learning-enabled natural-product purification

Priority	Proposed minimum test	Decision-relevant endpoint	Failure that should be visible
External chemical validation	Independent NP chemical classes	Retention/selectivity transfer	Domain-dependent error
Chromatographic transfer	New columns/modes/instruments	Method robustness	Method-specific degradation
Matrix challenge	Crude extracts and enriched fractions	Peak/fingerprint stability	Interference or missing features
Preparative validation	Increased loading and fraction collection	Purity, recovery, yield	Analytical-to-scale failure
Closed-loop comparison	Fixed experimental budget	Efficiency versus expert baseline	Objective/reward sensitivity
Interlaboratory validation	Independent sites and workflows	Reproducibility	Laboratory-specific model behavior
Mechanistic restraint	Independent physicochemical testing	Explanation fidelity	Interpretable association misread as mechanism

Note: The agenda is a proposed synthesis derived from the methodological gaps identified in this review; it is not a validated reporting standard.

CONCLUSION

Machine learning is changing natural-product separation through several distinct analytical functions rather than through one unified form of automated purification. Molecular models can predict retention and selectivity; signal models can detect or deconvolve chromatographic features; metabolomic representations can prioritize

fractions before complete structural characterization; and adaptive algorithms can guide method-development experiments. Their evidential meanings are different.

The central methodological contribution of this review is therefore to separate the object learned, the decision supported, and the validation level required. Analytical prediction should not be reported as preparative success, detected peaks should not be equated with purified

compounds, and fraction-level bioactivity prediction should not be converted into mechanistic certainty. The most consequential research gap lies at the analytical-to-preparative boundary, where natural-product matrices introduce loading, recovery, purity, yield, and transferability requirements poorly captured by existing benchmarks. Machine learning is consequently best positioned as a means of concentrating chromatographic experimentation and formalizing repeatable decisions while leaving chemical verification, failure diagnosis, and scale-dependent judgment explicitly within the purification workflow.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] Atanasov AG, Zotchev SB, Dirsch VM, Orhan IE, Banach M, Rollinger JM, et al. Natural products in drug discovery: Advances and opportunities. *Nat Rev Drug Discov.* 2021;20(3):200-16. doi:10.1038/s41573-020-00114-z
- [2] Nothias LF, Nothias-Esposito M, da Silva R, Wang M, Protsyuk I, Zhang Z, et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. *J Nat Prod.* 2018;81(4):758-67. doi:10.1021/acs.jnatprod.7b00737
- [3] Vitale GA, Geibel C, Minda V, Wang M, Aron AT, Petras D. Connecting metabolome and phenotype: Recent advances in functional metabolomics tools for the identification of bioactive natural products. *Nat Prod Rep.* 2024;41(6):885-904. doi:10.1039/D3NP00050H
- [4] Bosten E, Chen K, Hellings M, Cabooter D. Artificial intelligence for method development in liquid chromatography. *Trends Analyt Chem.* 2025;192:118320. doi:10.1016/j.trac.2025.118320
- [5] Liu Y, Yoshizawa AC, Ling Y, Okuda S. Insights into predicting small molecule retention times in liquid chromatography using deep learning. *J Cheminform.* 2024;16(1):113. doi:10.1186/s13321-024-00905-1
- [6] Domingo-Almenara X, Guijas C, Billings E, Montenegro-Burke JR, Uritboonthai W, Aisporna AE, et al. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nat Commun.* 2019;10(1):5811. doi:10.1038/s41467-019-13680-7
- [7] Kretschmer F, Harrieder EM, Hoffmann MA, Böcker S, Witting M. RepoRT: A comprehensive repository for small molecule retention times. *Nat Methods.* 2024;21(2):153-5. doi:10.1038/s41592-023-02143-z
- [8] Yang Q, Ji H, Lu H, Zhang Z. Prediction of liquid chromatographic retention time with graph neural networks to assist in small molecule identification. *Anal Chem.* 2021;93(4):2200-6. doi:10.1021/acs.analchem.0c04071
- [9] Kensert A, Bouwmeester R, Efthymiadis K, Van Broeck P, Desmet G, Cabooter D. Graph convolutional networks for improved prediction and interpretability of chromatographic retention data. *Anal Chem.* 2021;93(47):15633-41. doi:10.1021/acs.analchem.1c02988
- [10] Xu H, Lin J, Zhang D, Mo F. Retention time prediction for chromatographic enantioseparation by quantile geometry-enhanced graph neural network. *Nat Commun.* 2023;14(1):3095. doi:10.1038/s41467-023-38853-3
- [11] Kwon Y, Kwon H, Han J, Kang M, Kim JY, Shin D, et al. Retention time prediction through learning from a small training data set with a pretrained graph neural network. *Anal Chem.* 2023;95(47):17273-83. doi:10.1021/acs.analchem.3c03177
- [12] Vik D, Pii D, Mudaliar C, Nørregaard-Madsen M, Kontijevskis A. Performance and robustness of small molecule retention time prediction with molecular graph neural networks in industrial drug discovery campaigns. *Sci Rep.* 2024;14(1):8733. doi:10.1038/s41598-024-59620-4
- [13] Stienstra CMK, Nazdrajić E, Hopkins WS. From reverse phase chromatography to HILIC: Graph transformers power method-independent machine learning of retention times. *Anal Chem.* 2025;97(8):4461-72. doi:10.1021/acs.analchem.4c05859
- [14] Souihi A, Mohai MP, Palm E, Malm L, Krueve A. MultiConditionRT: Predicting liquid chromatography retention time for emerging contaminants for a wide range of eluent compositions and stationary phases. *J Chromatogr A.* 2022;1666:462867. doi:10.1016/j.chroma.2022.462867
- [15] Xu H, Lin J, Liu Q, Chen Y, Zhang J, Yang Y, et al. High-throughput discovery of chemical structure-polarity relationships combining automation and machine-learning techniques. *Chem.* 2022;8(12):3202-14. doi:10.1016/j.chempr.2022.08.008
- [16] Wu W, Xu H, Xu Y, Luo P, Zeng Q, Chen Y, et al. Intelligent column chromatography prediction model

- based on automation and machine learning. *Chem.* 2025;11(11):102598. doi:10.1016/j.chempr.2025.102598
- [17] Xu H, Wu W, Chen Y, Zhang D, Mo F. Explicit relation between thin film chromatography and column chromatography conditions from statistics and machine learning. *Nat Commun.* 2025;16(1):832. doi:10.1038/s41467-025-56136-x
- [18] Bosten E, Van Broeck P, Cabooter D. Automated tuning of denoising algorithms for noise removal in chromatograms. *J Chromatogr A.* 2023;1709:464360. doi:10.1016/j.chroma.2023.464360
- [19] Kensert A, Bosten E, Collaerts G, Efthymiadis K, Van Broeck P, Desmet G, et al. Convolutional neural network for automated peak detection in reversed-phase liquid chromatography. *J Chromatogr A.* 2022;1672:463005. doi:10.1016/j.chroma.2022.463005
- [20] Bueschl C, Doppler M, Varga E, Seidl B, Flasch M, Warth B, et al. PeakBot: Machine-learning-based chromatographic peak picking. *Bioinformatics.* 2022;38(13):3422-8. doi:10.1093/bioinformatics/btac344
- [21] Melnikov AD, Tsentalovich YP, Yanshole VV. Deep learning for the precise peak detection in high-resolution LC-MS data. *Anal Chem.* 2020;92(1):588-92. doi:10.1021/acs.analchem.9b04811
- [22] Zhang Z, Yang H, Wang Y, Zhang L, Lin SH. QuanFormer: A transformer-based precise peak detection and quantification tool in LC-MS-based metabolomics. *Anal Chem.* 2025;97(5):2698-706. doi:10.1021/acs.analchem.4c04531
- [23] Guo Z, Fan Y, Yu C, Lu H, Zhang Z. GCMSFormer: A fully automatic method for the resolution of overlapping peaks in gas chromatography-mass spectrometry. *Anal Chem.* 2024;96(15):5878-86. doi:10.1021/acs.analchem.3c05772
- [24] Gloaguen Y, Kirwan JA, Beule D. Deep learning-assisted peak curation for large-scale LC-MS metabolomics. *Anal Chem.* 2022;94(12):4930-7. doi:10.1021/acs.analchem.1c02220
- [25] Brittin NJ, Anderson JM, Braun DR, Rajski SR, Currie CR, Bugni TS. Machine learning-based bioactivity classification of natural products using LC-MS/MS metabolomics. *J Nat Prod.* 2025;88(2):361-72. doi:10.1021/acs.jnatprod.4c01123
- [26] Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, et al. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods.* 2020;17(9):905-8. doi:10.1038/s41592-020-0933-6
- [27] Bazzano CF, de Felicio R, Alves LFG, Costa JH, Ortega R, Vieira BD, et al. NP3 MS workflow: An open-source software system to empower natural product-based drug discovery using untargeted metabolomics. *Anal Chem.* 2024;96(19):7460-9. doi:10.1021/acs.analchem.3c05829
- [28] Bosten E, Pardon M, Chen K, Koppen V, Van Herck G, Hellings M, et al. Assisted active learning for model-based method development in liquid chromatography. *Anal Chem.* 2024;96(33):13699-709. doi:10.1021/acs.analchem.4c02700
- [29] Boelrijk J, Ensing B, Forré P, Pirok BWJ. Closed-loop automatic gradient design for liquid chromatography using Bayesian optimization. *Anal Chim Acta.* 2023;1242:340789. doi:10.1016/j.aca.2023.340789
- [30] Molenaar SRA, Bos TS, Boelrijk J, Dahlseid TA, Stoll DR, Pirok BWJ. Computer-driven optimization of complex gradients in comprehensive two-dimensional liquid chromatography. *J Chromatogr A.* 2023;1707:464306. doi:10.1016/j.chroma.2023.464306
- [31] Boelrijk J, Molenaar SRA, Bos TS, Dahlseid TA, Ensing B, Stoll DR, et al. Enhancing LC×LC separations through multi-task Bayesian optimization. *J Chromatogr A.* 2024;1726:464941. doi:10.1016/j.chroma.2024.464941
- [32] van Henten GB, Boelrijk J, Kattenberg C, Bos TS, Ensing B, Forré P, et al. Comparison of optimization algorithms for automated method development of gradient profiles. *J Chromatogr A.* 2025;1742:465626. doi:10.1016/j.chroma.2024.465626
- [33] Niezen LE, Libin PJK, Cabooter D, Desmet G. Reinforcement learning for automated method development in liquid chromatography: Insights in the reward scheme and experimental budget selection. *J Chromatogr A.* 2025;1748:465845. doi:10.1016/j.chroma.2025.465845
- [34] Subraveti SG, Li Z, Prasad V, Rajendran A. Can a computer “learn” nonlinear chromatography? Physics-based deep neural networks for simulation and optimization of chromatographic processes. *J Chromatogr A.* 2022;1672:463037. doi:10.1016/j.chroma.2022.463037