



Reaction Prediction Can Change How Natural Product Scaffolds Are Optimized by Connecting Biosynthetic Logic, Semisynthesis, and Medicinal Chemistry Search

Andrei Popescu^{1*}, Mihai Ionescu¹, Elena Stan², Cristina Radu³

¹Department of Reaction Prediction for Natural Products, Faculty of Pharmacy, University of Bucharest, Bucharest, Romania.

²Department of Biosynthetic Logic and Semisynthesis, Faculty of Pharmacy, University of Agricultural Sciences Cluj-Napoca, Cluj-Napoca, Romania.

³Department of Medicinal Chemistry Search and Scaffold Optimization, Faculty of Pharmacy, Polytechnic University of Bucharest, Bucharest, Romania.

ABSTRACT

Natural products offer chemically differentiated scaffolds, but their optimization is often framed as a problem of choosing desirable analogues rather than identifying the reactions that make those analogues reachable. This distinction matters because biosynthetic history, pre-existing stereochemical complexity, functional-group density, site selectivity, protecting-group requirements, and route context can sharply restrict which structural changes are experimentally plausible. This article develops an original methodological framework in which reaction prediction is positioned between natural-product scaffold representation and analogue prioritization. The framework integrates three otherwise partly separated sources of chemical knowledge: biosynthetic transformations as context-bearing priors, semisynthetic reactions that exploit accumulated scaffold complexity, and medicinal-chemistry reaction data that provide broader precedent for synthetic modification. It argues that predicted reactions should be evaluated not only by model likelihood but also by stereochemical and chemoselective compatibility, condition plausibility, route burden, analogue diversity, and downstream pharmacological value. Multistep routes, model failure, negative reaction evidence, and property prediction are treated as distinct analytical layers rather than collapsed into a single synthetic-accessibility score.

The principal contribution is a proposed reaction-aware search architecture for moving from an available natural-product scaffold to evidence-qualified analogue families. The framework is not presented as prospectively validated, and its usefulness will depend on scaffold-specific data coverage, representation quality, uncertainty calibration, route-level assessment, and experimental testing. Its purpose is to redirect optimization from asking which structures are attractive to asking which chemically meaningful transformations can make valuable chemical space reachable.

Key Words: Natural products, Reaction prediction, Semisynthesis, Biosynthetic transformations, Retrosynthesis, Analogue optimization

eIJPPR 2026; 16(3):147-155

HOW TO CITE THIS ARTICLE: Popescu A, Ionescu M, Stan E, Radu C. Reaction Prediction Can Change How Natural Product Scaffolds Are Optimized by Connecting Biosynthetic Logic, Semisynthesis, and Medicinal Chemistry Search. Int J Pharm Phytopharmacol Res. 2026;16(3):147-55. <https://doi.org/10.51847/fkzzC6Pd5l>

INTRODUCTION

Natural products remain important sources of pharmacologically differentiated structures because their architectures often occupy regions of chemical space that

are difficult to reproduce through conventional medicinal-chemistry libraries [1]. Yet optimization usually begins after a scaffold has already accumulated stereocenters, ring systems, oxidation patterns, and densely arranged functional groups. Under those conditions, the medicinal

Corresponding author: Andrei Popescu

Address: Department of Reaction Prediction for Natural Products, Faculty of Pharmacy, University of Bucharest, Bucharest, Romania.

E-mail: ✉ andrei.popescu@usamv.ro

Received: 21 February 2026; **Revised:** 10 May 2026; **Accepted:** 14 May 2026

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.



question is not simply which analogue would be desirable. It is which structural changes can be reached from the available scaffold without destroying the features that make the scaffold chemically or biologically interesting. Modern natural-product synthesis emphasizes that access itself shapes what can be investigated, while late-stage diversification demonstrates that existing complexity can be used rather than rebuilt when suitable reactions are available [2, 3]. These observations support a reaction-space formulation: an analogue is actionable only if at least one chemically credible transformation, or connected sequence of transformations, links it to a material starting point. Reaction space therefore differs from ordinary molecular similarity space. Two close analogues may have very different accessibility because one modification is chemoselective on the native scaffold and the other requires protecting groups, redox reorganization, stereochemical inversion, or an impractical sequence. Artificial-intelligence-assisted retrosynthesis has become increasingly relevant to medicinal chemistry because it can connect target structures with candidate disconnections and routes [4]. Natural-product optimization, however, presents a narrower and more constrained problem than generic target synthesis. The starting scaffold is often fixed; preservation of stereochemical information matters; many transformations are competing rather than independent; and the value of a reaction depends on what analogue family it opens, not only whether a single product can be proposed. This article therefore treats natural-product optimization as a reaction-aware search problem. The proposed methodological position is that reaction prediction should operate between scaffold representation and analogue prioritization, while remaining distinct from experimental feasibility. Biosynthetic precedent can inform which transformations are chemically coherent, semisynthesis can define the locally reachable neighborhood around the scaffold, and medicinal-chemistry data can expand the repertoire of possible modifications. Their integration is proposed here as a search architecture, not as evidence that a single model already performs the complete task.

Biosynthesis already encodes chemically plausible transformations

Biosynthesis provides more than historical explanation for natural-product structure. It records transformations that living systems use to assemble, oxidize, reduce, cyclize, transfer, and otherwise remodel chemically complex intermediates. Natural-product-specific retrosynthesis systems demonstrate that these transformations can be treated as learnable pathway objects rather than only as descriptive biochemical narratives [5, 6]. For reaction-aware optimization, that makes biosynthetic information potentially useful before any analogue is ranked: it can identify transformation families that are coherent with the scaffold's structural ancestry.

The value of biosynthetic precedent depends on representation. RAICHU shows that biosynthetic pathways can be formalized through explicit reaction rules and structurally resolved transformation sequences [7]. Such representation makes it possible to preserve the difference between a transformation class, the substrate context in which it occurs, and the pathway position that gives it meaning. A reaction-aware system should therefore retain precursor-product relationships and pathway context rather than compress biosynthetic knowledge into an undifferentiated list of reaction names.

Biochemical plausibility nevertheless does not equal synthetic practicality. Bioinspired total synthesis shows that biosynthetic reasoning can motivate laboratory transformations, but laboratory routes adapt that logic to reagents, conditions, selectivity requirements, and operational constraints that differ from enzyme-controlled environments [8]. The methodological implication is a boundary, not an equivalence: biosynthetic transformations can serve as priors that increase the chemical coherence of search, but they should not be treated as validated semisynthetic instructions.

The method components implied by this distinction are summarized in **Table 1**. The table separates what a reaction-aware system can inherit from biosynthetic evidence from what still requires synthetic interpretation.

Table 1. Method components for converting biosynthetic transformation knowledge into reaction-space priors

Method component	Computational object retained	Methodological use	Required boundary
Transformation precedent	Substrate-to-product change	Prioritize chemically coherent reaction families	Precedent is not synthetic feasibility
Pathway context	Ordered precursor-product relationships	Preserve where a transformation occurs in scaffold construction	Pathway position does not determine optimal laboratory order
Explicit reaction rule	Atom-level structural change	Support searchable and comparable transformation classes	Rule coverage is limited to encoded chemistry
Precedent retrieval	Related transformations and pathway examples	Condition predictions on structurally relevant prior cases	Retrieval quality depends on available records
Bioinspired translation	Biosynthetically motivated laboratory transformation	Connect biological logic to synthetic design	Enzymatic plausibility and reagent-based practicality remain distinct

Note: Entries describe methodological roles derived from the biosynthesis argument. They do not assign probabilities, yields, or experimental feasibility.

Semisynthesis defines a different search space from de novo synthesis

Semisynthesis begins from an experimentally available scaffold whose complexity has already been assembled. Its search space is therefore asymmetric: transformations are evaluated relative to a particular starting structure rather than to the abstract possibility of constructing any target molecule. High-throughput late-stage diversification illustrates this distinction by screening reactions directly against complex drug-like substrates and learning which local modifications are accessible within a defined experimental domain [9]. For natural products, the practical opportunity is similar. A successful transformation does not merely produce one analogue; it may reveal a repeatable entry point into an analogue family while preserving the scaffold features that would be costly to reconstruct.

This local search space is also strongly site dependent. Experimental C-H functionalization of terpene frameworks shows that apparently small changes can require careful control over which chemically similar position reacts and which remains untouched [10]. Such cases caution against representing semisynthetic accessibility as a simple function of molecular similarity. The distance between scaffold and analogue may be small in graph space yet large in reaction space if the desired site is masked by competing reactivity, stereochemical environment, or functional-group incompatibility.

The proposed framework therefore treats semisynthetic search as a constrained neighborhood around the available scaffold. Candidate reactions should be attached to explicit reactive sites and should carry information about what structural features must be retained, altered, or temporarily protected. De novo retrosynthesis remains valuable when the desired analogue lies outside this neighborhood, but it answers a different question: how to construct a target from available building blocks. Semisynthetic planning instead asks which transformations can exploit complexity that is already present. Keeping these search modes separate prevents a model from rewarding theoretically constructible analogues that are poorly matched to the practical advantage of starting from the natural product itself.

Reaction prediction as a bridge between scaffold and analog

Reaction prediction provides a computational bridge between an available scaffold and candidate structural changes. Early machine-learning approaches established that organic reaction outcomes can be learned from reaction corpora, while the Molecular Transformer showed that sequence-based prediction can additionally represent uncertainty in predicted products [11, 12]. These

capabilities are relevant to semisynthetic planning because they allow the search object to shift from “which analogue is desirable?” to “which transformation on this scaffold is sufficiently plausible to justify further consideration?” The distinction is methodological: predicted likelihood proposes a transition in chemical space; it does not establish laboratory feasibility.

How that transition is represented matters. Iterative reaction-edit models demonstrate that retrosynthesis can be formulated through explicit structural edits rather than only unconstrained generation of complete precursor strings [13]. For complex natural products, this suggests an advantage in maintaining the reaction center, retained scaffold, and altered atoms as separable information. Such representations could make it easier to enforce site conservation and later attach stereochemical, chemoselective, and route-context constraints.

Prediction quality also depends on the reaction records from which models learn. ORDERly demonstrates that curation, standardization, and benchmark construction materially affect the data used to train and compare reaction models [14]. A reaction-aware natural-product method should therefore carry data provenance alongside the predicted transformation. A high-ranked reaction learned from abundant medicinal-chemistry precedent and a low-support prediction extrapolated to an unusual polycyclic scaffold should not be treated as evidentially equivalent.

The bridge proposed here is consequently conditional. Reaction prediction expands the set of candidate scaffold-to-analogue transitions, but subsequent layers must test whether those transitions preserve stereochemistry, respect competing functionality, admit plausible conditions, and remain useful within a multistep route. Prediction generates candidates for chemical reasoning; it should not replace that reasoning.

Representing stereochemical and chemoselective constraints

Natural-product reaction prediction becomes unreliable if a scaffold is represented only as connectivity plus a desired bond change. Stereochemical state, conformationally biased reactive sites, neighboring functionality, and competing reaction centers can determine whether a nominally reasonable transformation gives the intended product. Chemistry-aware machine learning for Diels-Alder reactions illustrates that incorporating chemically meaningful structure can improve outcome prediction in a selectivity-sensitive reaction class [15]. The result does not establish transferability to natural products, but it supports the broader methodological requirement that reaction representations preserve information relevant to selectivity

rather than treating all formally valid transformations as equivalent.

Reaction context also includes conditions. A two-stage neural model for reaction-condition prediction demonstrates that feasible-condition selection can be modeled as a task distinct from product prediction [16]. For semisynthetic planning, this separation is important because a transformation compatible with one functional group may be incompatible with another under the required reagent, catalyst, solvent, temperature, or redox environment. Condition plausibility should therefore enter as a constraint on a predicted transformation, not as an afterthought attached to an otherwise completed route.

Protecting-group requirements and route context add another layer. Protection-aware synthesis planning shows that routes judged attractive without protection logic can change when temporary functional-group management and route scoring are considered [17]. In densely functionalized natural products, that burden may determine whether a local edit remains advantageous relative to an alternative sequence.

The proposed representation is consequently multicomponent: a candidate transformation should retain the reactive site, stereochemical state, competing functional groups, predicted conditions, and any protection or deprotection dependencies. These fields are not claimed to guarantee success. They are intended to prevent reaction likelihood from being mistaken for scaffold-compatible chemistry and to provide explicit reasons for rejecting a superficially plausible analogue path before costly experimental testing.

Learning from medicinal chemistry reaction data without losing natural product context

Medicinal-chemistry reaction corpora offer scale, but scale does not guarantee that the learned chemistry reflects the structural regime of natural products. Quantitative interpretation of reaction-prediction models has shown that apparently strong performance can coexist with learned biases in the underlying datasets [18]. For natural-product optimization, the concern is whether such shortcuts are poorly aligned with stereochemically dense, polycyclic, or otherwise uncommon scaffolds. Training-domain proximity should therefore qualify each prediction.

The absence of unsuccessful chemistry creates a second problem. Negative reaction data can improve reaction-outcome prediction because it teaches models to discriminate between transformations that were attempted and failed and those that produced the intended chemistry

[19]. For natural-product semisynthesis, an unreported transformation may be unattempted, failed, or simply unpublished. These states cannot be collapsed. The proposed framework therefore separates documented negative evidence from absence of precedent.

Evaluation must also move beyond a single top-k success statistic. Granular retrosynthesis metrics show that predictions can fail in chemically different ways even when aggregate performance appears similar [20]. For natural-product use, those failure classes should include wrong site, wrong stereochemical consequence, incompatible functional-group context, missing protection logic, and route-level dead ends. These categories are proposed as application-specific diagnostics rather than as established frequencies. Their purpose is to make model failure chemically interpretable before a reaction is elevated into an analogue-design recommendation.

Biosynthetic transformations as priors for reaction prediction

Biosynthetic transformations should enter reaction prediction as a distinct source of prior information rather than being merged indiscriminately with general synthetic reaction data. Enzymatic retrosynthesis requires domain-specialized modeling because enzyme-mediated transformations encode substrate context and transformation patterns that differ from conventional reaction corpora [21]. This supports keeping biosynthetic or enzymatic evidence identifiable inside the model, even when a common scaffold representation is used.

Hybrid planning systems further show that organic and enzymatic reaction spaces can be searched together computationally [22]. For natural products, the scaffold can serve as the common state connecting those spaces. A biosynthetic prior may raise the plausibility of a transformation family; semisynthetic evidence may indicate whether related chemistry is achievable on a complex scaffold; medicinal-chemistry precedent may suggest reagents, protecting strategies, or alternative transformations. None of these evidence types should automatically override the others.

The proposed workflow therefore treats prior information as conditional rather than deterministic. Biosynthetic precedent can expand or reweight candidate reactions, but predicted transformations must still survive scaffold-specific stereochemical, chemoselective, condition, and route-context checks. **Table 2** organizes the resulting workflow as a sequence of decisions rather than a single prediction score.

Table 2. Proposed workflow for converting reaction precedent into natural-product analogue candidates

Workflow step	Input retained	Decision operation	Output passed forward
Scaffold state	Structure, stereochemistry, functional groups	Define modifiable and protected features	Reaction-ready scaffold representation
Prior assembly	Biosynthetic, semisynthetic, medicinal-chemistry precedent	Keep evidence sources distinguishable	Context-tagged reaction candidates
Reaction proposal	Candidate transformations and sites	Predict plausible scaffold-to-product changes	Ranked local transformations
Constraint checking	Site, stereo, chemoselectivity, conditions, protection needs	Reject context-incompatible proposals	Scaffold-compatible candidates
Route expansion	Surviving reactions and intermediates	Search connected one- or multistep paths	Reachable analogue families
Value ranking	Accessibility, diversity, predicted properties	Compare candidates without conflating objectives	Experiment-ready priorities
Validation boundary	Route rationale and uncertainty	Require expert review and prospective testing	Evidence-qualified decisions

The full synthesis is shown in **Figure 1**. The figure should be read as a reaction-space search architecture in which multiple evidence sources condition candidate

transformations, while route failure and experimental validation remain outside the predictive core.

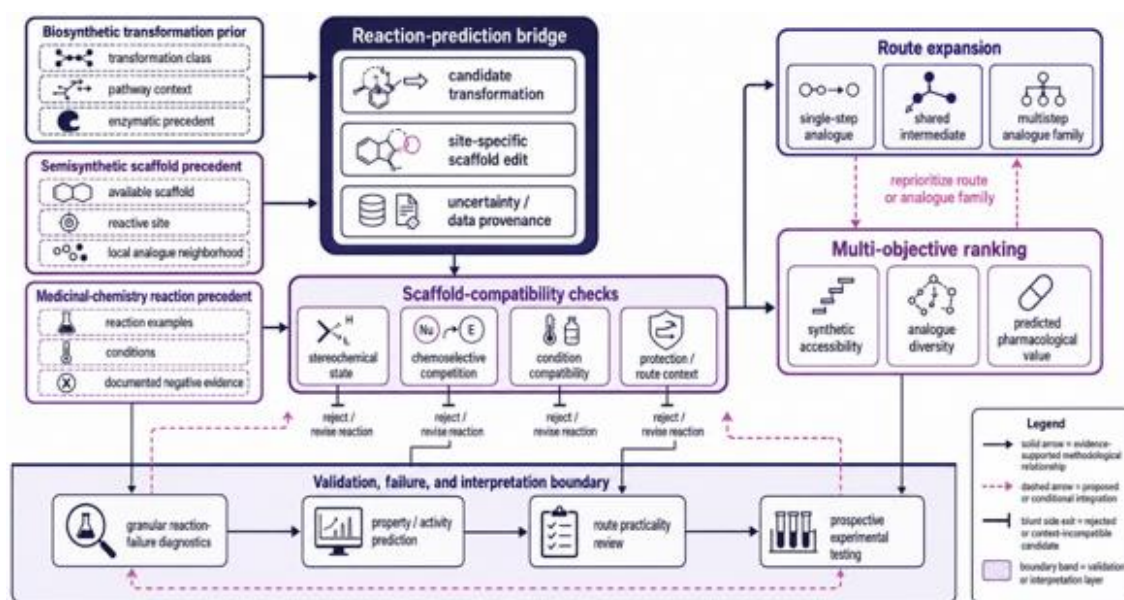


Figure 1. From natural-product scaffold state to evidence-qualified reaction-space search: a proposed reaction-aware optimization framework

A biosynthesis–semisynthesis–medicinal chemistry search framework

The proposed framework treats a natural-product optimization campaign as a search over chemically reachable states. Route-search systems such as AiZynthFinder demonstrate that single-step predictions become useful only when embedded in explicit search over connected transformations [23]. Hypergraph exploration similarly shows that multistep retrosynthesis requires managing alternative branches and dependencies rather than ranking each reaction independently [24]. For semisynthesis, a locally attractive transformation should also be evaluated by what it makes reachable next. The scaffold representation also needs to preserve the nature of the structural change. Molecular-assembly

approaches show that retrosynthetic decisions can be expressed through explicit structural operations [25]. Pathway-aware retrosynthesis further indicates that surrounding route context can influence which disconnection is preferred [26]. These findings motivate a proposed state representation carrying the retained scaffold, transformed site, stereochemical state, route history, protection status, and evidence provenance. Candidate reactions must then be reranked according to their actual transformation logic. RetroRanker demonstrates that reaction-change information can improve candidate ordering [27]. The framework extends that idea by proposing three linked but non-equivalent ranking dimensions: accessibility, the structural diversity opened by the reaction, and downstream pharmacological

value. A modest-probability reaction may deserve attention if it opens a distinct analogue family, whereas a high-probability reaction may be strategically weak if it yields redundant analogues or creates route burden.

This architecture is model-agnostic and does not require a particular transformer, graph network, template system, or search algorithm. Its methodological claim is narrower: natural-product optimization should preserve reaction identity, scaffold context, and route consequences long enough for them to influence analogue selection.

Ranking reactions by accessibility, diversity, and pharmacological value

Accessibility should be treated as multidimensional. Reaction likelihood, condition plausibility, protection burden, number and topology of downstream steps, and dependence on uncommon intermediates describe different sources of synthetic risk. They should not be collapsed into a universal scalar without empirical calibration. The framework instead proposes an explicit multi-criterion comparison in which each component remains inspectable. The search object may also be an entire path rather than a single transformation. DirectMultiStep shows that multistep retrosynthetic routes can be generated as connected objects rather than reconstructed only through repeated independent single-step calls [28]. This supports evaluating analogue families through the routes that jointly produce them. If several analogues share an intermediate, the synthetic value of reaching that intermediate may exceed the value of any one final analogue considered alone.

Route topology adds another criterion. Convergent retrosynthesis planning demonstrates that route development can be altered by objectives concerning how branches join and how synthesis is organized [29]. For natural-product semisynthesis, convergence will not always be relevant, but route structure still matters when parallel derivatization, common intermediates, or late diversification can reduce repeated synthetic effort.

Pharmacological value should enter only after this chemical structure is preserved. Predicted activity or developability can help prioritize reachable analogues, but it should not retroactively convert a chemically weak route into a feasible one. The proposed ranking therefore asks three sequentially connected questions: can the reaction plausibly be executed on this scaffold, what distinct analogue space does it open, and is that reachable space worth testing biologically?

Multi-step paths to analog families

Natural-product optimization is often more efficient when a route is evaluated for the family of analogues it can

support rather than for a single terminal structure. A common intermediate may permit several late-stage modifications. The proposed framework therefore treats intermediate states as reusable search assets and records the analogue branches they enable.

Multistep prediction introduces its own uncertainty. Evaluation of transformer-based retrosynthesis shows that satisfactory behavior at the single-step level does not automatically translate into satisfactory multistep planning [30]. Local errors can accumulate, search may propagate an implausible intermediate, and route ranking can amplify early mistakes. Natural-product scaffolds add further risks because one incorrect stereochemical or site assignment may invalidate every downstream branch.

For this reason, path confidence should not be calculated as though successive predictions were independent guarantees. The framework instead proposes a route-risk record that identifies the weakest transformation, unsupported scaffold context, unresolved protection requirement, and any step that relies on extrapolation beyond close precedent. This makes a family-level path useful for decision making without implying that the entire route has been validated.

Reaction prediction failure in structurally complex scaffolds

Complex natural products create failure modes that ordinary reaction benchmarks may underrepresent. A model may propose the correct reaction class at the wrong site, preserve connectivity but lose stereochemical information, ignore competing functionality, or favor a disconnection that becomes impractical once protection and route context are considered. These are chemically distinct failures and should remain distinguishable during evaluation.

Molecular-property-aware retrosynthesis demonstrates that auxiliary molecular information can be incorporated into reaction prediction [31]. That supports, but does not prove, richer scaffold representations for natural products. The proposed extension is to treat scaffold state as more than a molecular graph: the state should include stereochemical commitments, sensitive functional groups, route history, and confidence in whether the prediction is interpolating from related chemistry or extrapolating beyond it.

Table 3 compares the proposed reaction-aware use case with common computational abstractions. The distinction is not that existing methods are inadequate in general; it is that natural-product semisynthesis requires additional context before a prediction can be interpreted as an optimization opportunity.

Table 3. Comparison of reaction-prediction abstractions for natural-product scaffold optimization

Methodological abstraction	Primary object optimized	Context usually emphasized	Natural-product risk if used alone	Proposed reaction-aware addition
Forward reaction prediction	Product from given reactants	Learned reaction precedent	Correct class may map to wrong site or stereochemical state	Scaffold-compatibility checks
Single-step retrosynthesis	Immediate precursors	Disconnection plausibility	Local success can ignore downstream burden	Route-history retention
Generic route search	Connected synthesis path	Search depth, templates, route score	Feasible-looking route may depend on poor scaffold context	Context-tagged path risk
Condition prediction	Reagents and reaction environment	Recorded successful conditions	Conditions may not transfer to dense functionality	Functional-group compatibility
Property-aware prediction	Molecular or auxiliary properties	Learned molecular features	Property gain may outrank inaccessible chemistry	Separate reachability and value
Proposed integrated search	Reachable analogue family	Reaction, scaffold, route, uncertainty, value	Still dependent on data coverage and model calibration	Prospective experimental rejection tests

Integration with property and activity prediction

Synthetic reachability and pharmacological desirability are different prediction problems. Multitask graph neural networks for ADME prediction illustrate how molecular properties can be estimated in lead-optimization settings [32], but those outputs do not establish that an analogue can be synthesized or that a predicted property will reproduce experimentally. The reaction-aware framework therefore keeps property and activity prediction downstream from, but connected to, chemical reachability. This separation enables a useful search strategy. Reaction prediction can define the set of plausibly reachable analogues; property or activity models can then rank within that set; and disagreement between chemical and pharmacological objectives can trigger reprioritization. An analogue with attractive predicted properties but a fragile route remains a high-uncertainty candidate, whereas a readily accessible analogue with modest predicted value may serve as an experimentally informative probe of structure-activity space.

The proposed framework does not prescribe a universal weighting between these objectives. Weighting should depend on campaign stage, assay confidence, material availability, and the cost of testing. What must remain invariant is the distinction between predicted chemical accessibility, predicted biological or developability value, and measured experimental performance.

Experimental validation and route practicality

A reaction proposal becomes decision-relevant only when its uncertainty is made explicit. Predicted conditions, route scores, and multistep plans remain computational evidence, not substitutes for experimental execution. For natural products, expert review remains important because substrate effects, reagent incompatibilities, purification, and material constraints may be absent from training data. Experimental route demonstrations accompanying site-specific retrosynthetic planning show the value of moving

beyond retrospective benchmark success toward laboratory testing [33]. Such demonstrations strengthen evidence for particular predictions, but their limited scope cannot establish general performance across natural-product chemical space. The proposed framework therefore treats prospective testing as a boundary condition rather than a final decorative validation step.

A practical validation sequence should begin with reaction-level interrogation: confirm the intended site, stereochemical outcome, and compatibility with neighboring functionality. It should then assess whether protection and deprotection operations preserve route advantage, whether intermediates are isolable, and whether the path can be reproduced under realistic material constraints. Failed reactions should be retained as data rather than discarded, because they can refine scaffold-specific priors and reveal where generic reaction models extrapolate poorly.

Accordingly, the framework is falsifiable at several levels. A predicted transformation can fail chemically; a route can prove operationally impractical; an analogue family can be reachable but pharmacologically uninformative; or model rankings can fail to outperform simpler chemist-defined choices. These outcomes would refine or challenge the proposed architecture rather than being treated as implementation noise.

CONCLUSION

Natural-product optimization should not begin from the assumption that every attractive analogue is an equally meaningful design target. Existing evidence separately supports biosynthetic-pathway modeling, late-stage diversification, reaction prediction, condition prediction, route search, multistep planning, and property prediction. The methodological opportunity is to connect these capabilities without erasing the boundaries between them.

The framework proposed here therefore places reaction space between scaffold and analogue. Biosynthetic transformations provide context-bearing priors; semisynthesis defines the locally reachable neighborhood; medicinal-chemistry data broaden reaction precedent; scaffold constraints filter candidate transformations; route search organizes connected paths; and property models help rank what remains chemically reachable. Failure analysis and prospective experimentation remain outside the predictive core because computational plausibility is not synthetic validation.

The framework is intended as a testable organization of the optimization problem, not as a claim of proven superiority. Its value will depend on natural-product-specific data coverage, faithful stereochemical and functional-group representation, treatment of negative evidence, uncertainty calibration, and experimental rejection of incorrect routes. The central shift is simple but consequential: optimize not only the molecules one would like to make, but the reactions that determine which chemical space can actually be reached.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

- [1] Atanasov AG, Zotchev SB, Dirsch VM, International Natural Product Sciences Taskforce, Supuran CT. 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] Shenvi RA. Natural product synthesis in the 21st century: Beyond the mountain top. *ACS Cent Sci.* 2024;10(3):519-28. doi:10.1021/acscentsci.3c01518
- [3] Hong B, Luo T, Lei X. Late-stage diversification of natural products. *ACS Cent Sci.* 2020;6(5):622-35. doi:10.1021/acscentsci.9b00916
- [4] Long L, Li R, Zhang J. Artificial intelligence in retrosynthesis prediction and its applications in medicinal chemistry. *J Med Chem.* 2025;68(3):2333-55. doi:10.1021/acs.jmedchem.4c02749
- [5] Zheng S, Zeng T, Li C, Chen B, Coley CW, Yang Y, et al. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nat Commun.* 2022;13(1):3342. doi:10.1038/s41467-022-30970-9
- [6] Kim T, Lee S, Kwak Y, Choi MS, Park J, Hwang SJ, et al. READRetro: Natural product biosynthesis predicting with retrieval-augmented dual-view retrosynthesis. *New Phytol.* 2024;243(6):2512-27. doi:10.1111/nph.20012
- [7] Terlouw BR, Biermann F, Vromans SPJM, Zamani E, Helfrich EJN, Medema MH. RAICHU: Automating the visualisation of natural product biosynthesis. *J Cheminform.* 2024;16(1):106. doi:10.1186/s13321-024-00898-x
- [8] Chen L, Chen P, Jia Y. Bioinspired total synthesis of natural products. *Acc Chem Res.* 2024;57(24):3524-40. doi:10.1021/acs.accounts.4c00654
- [9] Nippa DF, Atz K, Hohler R, Müller AT, Marx A, Bartelmus C, et al. Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *Nat Chem.* 2024;16(2):239-48. doi:10.1038/s41557-023-01360-5
- [10] Kang YC, Wetterer RT, Karimov RR, Kojima M, Surke M, Martín-Torres I, et al. Substitution, elimination, and integration of methyl groups in terpenes initiated by C-H bond functionalization. *ACS Cent Sci.* 2024;10(11):2016-27. doi:10.1021/acscentsci.4c01108
- [11] Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. Prediction of organic reaction outcomes using machine learning. *ACS Cent Sci.* 2017;3(5):434-43. doi:10.1021/acscentsci.7b00064
- [12] 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
- [13] Han Y, Xu X, Hsieh CY, Ding K, Xu H, Xu R, et al. Retrosynthesis prediction with an iterative string editing model. *Nat Commun.* 2024;15(1):6404. doi:10.1038/s41467-024-50617-1
- [14] Wigh DS, Arrowsmith J, Pomberger A, Felton KC, Lapkin AA. ORDERly: Data sets and benchmarks for chemical reaction data. *J Chem Inf Model.* 2024;64(9):3790-8. doi:10.1021/acs.jcim.4c00292
- [15] Keto A, Guo T, Underdue M, Stuyver T, Coley CW, Zhang X, et al. Data-efficient, chemistry-aware machine learning predictions of Diels-Alder reaction outcomes. *J Am Chem Soc.* 2024;146(23):16052-61. doi:10.1021/jacs.4c03131
- [16] Chen LY, Li YP. Enhancing chemical synthesis: A two-stage deep neural network for predicting feasible reaction conditions. *J Cheminform.* 2024;16(1):11. doi:10.1186/s13321-024-00805-4
- [17] Westerlund AM, Sigmund LM, Mijangos MV, Kannas C, Genheden S, Kabeshov M. Toward lab-ready AI synthesis plans with protection strategies and



- route scoring. *J Chem Inf Model.* 2026;66(11):6361-75. doi:10.1021/acs.jcim.6c01147
- [18] Kovács DP, McCorkindale W, Lee AA. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nat Commun.* 2021;12(1):1695. doi:10.1038/s41467-021-21895-w
- [19] Toniato A, Vaucher AC, Laino T, Graziani M. Negative chemical data boosts language models in reaction outcome prediction. *Sci Adv.* 2025;11(24). doi:10.1126/sciadv.adt5578
- [20] Tadanki AS, Rao HSP, Priyakumar UD. Dissecting errors in machine learning for retrosynthesis: A granular metric framework and a transformer-based model for more informative predictions. *Digit Discov.* 2025;4(3):831-45. doi:10.1039/D4DD000263F
- [21] Liu T, Xu J, Zhan X, Lin S, Siu SWI. Enzyformer: A two-stage pretrained model for enzymatic retrosynthesis. *J Cheminform.* 2026;18(1):35. doi:10.1186/s13321-026-01164-y
- [22] Wang X, Yin X, Zhang X, Zhao H, Gu S, Wu Z, et al. A virtual platform for automated hybrid organic-enzymatic synthesis planning. *Nat Commun.* 2025;16(1):10929. doi:10.1038/s41467-025-65898-3
- [23] Genheden S, Thakkar A, Chadimová V, Reymond JL, Engkvist O, Bjerrum E. AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. *J Cheminform.* 2020;12(1):70. doi:10.1186/s13321-020-00472-1
- [24] Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem Sci.* 2020;11(12):3316-25. doi:10.1039/C9SC05704H
- [25] Wang Y, Pang C, Wang Y, Jin J, Zhang J, Zeng X, et al. Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks. *Nat Commun.* 2023;14(1):6155. doi:10.1038/s41467-023-41698-5
- [26] Zhang JJ, Ha SK, Roh J, Tu Z, Verma P, Coley CW, et al. Pathway-aware template-based retrosynthesis. *J Chem Inf Model.* 2026;66(13):7489-500. doi:10.1021/acs.jcim.6c01458
- [27] Li J, Fang L, Lou JG. RetroRanker: Leveraging reaction changes to improve retrosynthesis prediction through re-ranking. *J Cheminform.* 2023;15(1):58. doi:10.1186/s13321-023-00727-7
- [28] Shee Y, Morgunov A, Li H, Batista VS. DirectMultiStep: Direct route generation for multistep retrosynthesis. *J Chem Inf Model.* 2025;65(8):3903-14. doi:10.1021/acs.jcim.4c01982
- [29] Torren-Peraire P, Verhoeven J, Herman D, Ceulemans H, Tetko IV, Wegner JK. Improving route development using convergent retrosynthesis planning. *J Cheminform.* 2025;17:26. doi:10.1186/s13321-025-00953-1
- [30] Westerlund AM, Manohar Koki S, Kancharla S, Tibo A, Saigiridharan L, Kabeshov M, et al. Do Chemformers dream of organic matter? Evaluating a transformer model for multistep retrosynthesis. *J Chem Inf Model.* 2024;64(8):3021-33. doi:10.1021/acs.jcim.3c01685
- [31] Liu M, Xiao F, Yu Z, Huang H, Li Y, Yang Y, et al. RetroMPA: A molecular property-aware auxiliary framework for enhancing retrosynthesis prediction. *J Chem Inf Model.* 2026. doi:10.1021/acs.jcim.6c01506
- [32] Ito S, Koyama T, Matsumoto S, Kojima R, Okamoto Y, Kuroda M, et al. Improving ADME prediction with multitask graph neural networks and assessing explainability in lead optimization. *J Chem Inf Model.* 2025;65(21):11796-803. doi:10.1021/acs.jcim.5c00758
- [33] Shee Y, Li H, Zhang P, Nikolic AM, Lu W, Kelly HR, et al. Site-specific template generative approach for retrosynthetic planning. *Nat Commun.* 2024;15(1):7818. doi:10.1038/s41467-024-52048-4