Safety-by-Design Generative Chemistry for Natural Products: Toxicity Boundaries, Interaction Risk, Validation Gates, and Human Oversight

Nguyen Thanh Huy[1]*, Pham Quang Minh[1], Le Thi Bich[2]

[1]Department of AI for Vietnamese Traditional Medicine (Thuốc Nam) Informatics, Faculty of Pharmacy, Hanoi University of Pharmacy, Hanoi, Vietnam

[2]Department of Computational Screening for Antiviral Lectins, Faculty of Pharmacy, Can Tho University of Medicine and Pharmacy, Can Tho, Vietnam.

ABSTRACT

Generative chemistry offers a means of exploring natural product-inspired chemical space beyond directly isolated compounds, but molecular novelty and predicted activity do not establish safety, developability, or therapeutic utility. This article proposes a safety-by-design framework for integrating safety considerations throughout natural product-inspired molecular generation rather than applying toxicity filters only after candidate production. The framework links research-objective definition, natural product scaffold selection, molecular generation, natural product-likeness assessment, bioactivity hypothesis screening, and multi-objective optimization with explicit toxicity boundaries, ADMET constraints, structural-alert assessment, reactive-group review, off-target evaluation, safety-pharmacology concerns, and metabolism-related risks. Herb–drug and drug–drug interaction risks are treated as independent evidence domains because intrinsic molecular toxicity does not capture risks arising from co-exposure, metabolic interference, transport modulation, or compound-mixture effects. Model reliability is addressed through uncertainty estimation, applicability-domain assessment, out-of-distribution detection, data provenance, bias review, and documentation. Sequential validation gates are proposed to distinguish computationally plausible candidates from hypotheses that warrant in vitro, toxicological, pharmacological, or, where appropriate, in vivo investigation. Expert medicinal chemistry, toxicology, safety pharmacology, and interaction-risk review remain necessary, with human override retained at every advancement decision. The principal contribution is a structured architecture that can reduce unsafe prioritization and improve early-stage decision discipline while maintaining strict no-safety-claim and no-clinical-claim boundaries.

Key Words: *Safety-by-design, Generative chemistry, Natural products, Toxicity prediction, Interaction risk, Validation gates*

eIJPPR 2026; 16(3):135-146

HOW TO CITE THIS ARTICLE: Huy NT, Minh PQ, Bich LT. Safety-by-Design Generative Chemistry for Natural Products: Toxicity Boundaries, Interaction Risk, Validation Gates, and Human Oversight. Int J Pharm Phytopharmacol Res. 2026;16(3):135-46. https://doi.org/10.51847/HjuyAMdFCC

INTRODUCTION

Artificial intelligence is increasingly used to support target identification, molecular-property prediction, candidate prioritization, and other stages of drug discovery, while generative molecular design extends these applications by proposing new chemical structures in relation to predefined objectives. Inverse molecular design can therefore be understood as a constrained search process in which molecular representations are explored against specified property goals rather than as an autonomous mechanism for discovering intrinsically suitable medicines [1, 2]. This distinction is particularly important in natural product-inspired discovery, where structurally distinctive scaffolds, stereochemical complexity, and biological precedent may create attractive starting points for

Corresponding author: Nguyen Thanh Huy
Address: Department of AI for Vietnamese Traditional Medicine (Thuốc Nam) Informatics, Faculty of Pharmacy, Hanoi University of Pharmacy, Hanoi, Vietnam.
E-mail: ✉ huy.nguyen@hup.edu.vn
Received: 02 March 2026; **Revised:** 19 June 2026; **Accepted:** 22 June 2026

molecular generation but do not remove the need for systematic safety evaluation.

Generative chemistry can expand the range of hypotheses derived from natural products by proposing analogues, scaffold variants, or property-conditioned candidates that would not be encountered through direct isolation or conventional analogue enumeration alone. Nevertheless, the practical interpretation of these outputs depends on iterative interaction among computational design, medicinal chemistry, synthesis feasibility, biological testing, and expert judgment [3]. A generated structure is therefore best described as a candidate hypothesis: a computational proposal whose identity, plausibility, activity, safety, and developability remain to be established through progressively stronger evidence.

Novelty, predicted target activity, and natural product-likeness are inadequate as stand-alone prioritization criteria. A structure may appear chemically valid and score favourably against an activity objective while containing unstable functionality, reactive groups, metabolic liabilities, off-target potential, interaction risks, or features that lie outside the evidentiary domain of the models used to assess it. Medicinal-chemistry value cannot be inferred from novelty or favourable computational scoring alone, because generated molecules may be impractical, chemically undesirable, or insufficiently supported by experimental evidence [4]. Natural product origin or resemblance likewise does not guarantee safety; many biologically active natural substances produce dose-dependent toxicity, pharmacokinetic interactions, or context-specific adverse effects.

This article advances an original safety-by-design framework for natural product-inspired generative chemistry. The proposed architecture embeds safety considerations into research-objective definition, scaffold selection, molecular generation, multi-objective optimization, toxicity and ADMET assessment, interaction-risk evaluation, model-reliability review, validation planning, expert oversight, and feedback updating. It distinguishes candidate generation from candidate validation and introduces explicit boundaries preventing computationally screened molecules from being described as safe, clinically useful, or development-ready. The framework is intended to support research prioritization by reducing avoidable unsafe advancement, clarifying evidence gaps, and directing generated hypotheses toward proportionate experimental evaluation.

Safety-by-design in generative chemistry

Safety-by-design in generative chemistry can be defined as the prospective integration of safety objectives, exclusion rules, reliability checks, evidence requirements, and human decision boundaries throughout molecular design. It differs from post hoc toxicity filtering because safety-relevant considerations are incorporated before, during, and after generation rather than being applied only to a final candidate list. De novo molecular design is inherently multi-objective: validity, novelty, intended activity, physicochemical properties, synthetic accessibility, developability, and safety may conflict, and improving one objective may degrade another [5]. Safety-by-design therefore does not seek a single composite "safety score." Instead, it requires explicit endpoint-level constraints, preserved uncertainty, and transparent handling of competing objectives.

The framework also differs from unconstrained generation, conventional virtual screening, and purely activity-driven optimization. Unconstrained generation prioritises chemical-space exploration without necessarily specifying acceptable liability boundaries. Conventional virtual screening generally ranks pre-existing structures against a target or property model, whereas generative systems actively construct new structures and may exploit weaknesses in objective functions. Activity-driven optimization can similarly reward potency-associated features while overlooking instability, reactivity, metabolic risk, or unrealistic synthesis. Synthetic accessibility must therefore be assessed independently because formally valid and highly scored generated structures may remain impractical to prepare or may receive misleadingly favourable feasibility estimates from simplified proxy metrics [6]. A safety-by-design workflow should treat inaccessible or chemically implausible structures as failed design hypotheses rather than merely lower-ranked candidates.

For natural product-inspired discovery, safety-by-design begins with verified natural product identity, scaffold provenance, stereochemical representation, and a clear explanation of what "inspiration" means in the specific design context. A natural product-inspired scaffold may retain a core topology, privileged three-dimensional arrangement, pharmacophoric relationship, or distributional resemblance without reproducing the source compound. Prospective work has shown that machine-assisted design can generate natural product-inspired molecules that proceed to synthesis and targeted biological evaluation, demonstrating the value of such systems for producing testable bioactivity hypotheses [7]. However, confirmation of activity against a selected target does not establish broad toxicological acceptability, interaction safety, exposure suitability, or clinical utility.

Natural product-likeness should consequently function as a design descriptor rather than a surrogate for safety or therapeutic value. Generative models can be conditioned toward structural features associated with natural products, expanding scaffold diversity while preserving selected

distributional characteristics [8]. Within a safety-by-design architecture, this capability should be coupled to novelty constraints, synthetic-accessibility review, toxicity and ADMET prediction, structural-alert screening, interaction-risk assessment, uncertainty estimation, and applicability-domain evaluation. Human experts must retain authority to redefine objectives, reject misleading outputs, request additional analyses, and prevent candidates from advancing when favourable model predictions are unsupported by relevant evidence.

Toxicity boundaries and interaction risk

A toxicity boundary is an evidence-qualified decision limit used to determine whether a generated candidate may proceed, requires redesign, should be deferred for additional evidence, or should be rejected from a particular workflow stage. It is not a universal threshold separating "safe" from "unsafe" chemistry. Safety screening requires endpoint-specific molecular benchmarks, multi-endpoint ADMET prediction, and interpretable structural-alert assessment because each provides a different form of evidence [9-11]. Benchmark performance indicates how models behave on defined datasets; ADMET models estimate distinct absorption, distribution, metabolism, excretion, and toxicity properties; and structural alerts identify motifs associated with possible liabilities. None of these evidence types independently establishes toxicological safety, and favourable predictions across several endpoints cannot eliminate unmeasured hazards.

Toxicity-boundary assessment should extend beyond intrinsic toxicity predictions to reactive-group screening, off-target risk, and safety-pharmacology concerns. Structural alerts or putative toxicophores may indicate electrophilicity, instability, promiscuity, or bioactivation potential, but their significance depends on molecular context, metabolism, concentration, and exposure. Conversely, the absence of recognised alerts cannot be interpreted as evidence of safety. Biological-signature approaches that compare compounds across chemical, target, network, cellular, and clinical information spaces may support hypotheses about unintended targets or mechanism-level similarities that are not visible from structure alone [12]. Such outputs should guide counterscreening and expert review, not be treated as demonstrated binding, causality, or organism-level toxicity.

Interaction risk requires a distinct assessment layer because a molecule with an apparently acceptable intrinsic toxicity profile may still alter the exposure or effects of another substance. Drug–drug and drug–food interaction models can identify plausible relations from molecular, pharmacological, or network evidence, supporting the prioritization of interaction mechanisms for further testing [13]. In natural product-inspired discovery, this layer should consider enzyme inhibition or induction, transporter modulation, pharmacodynamic overlap, metabolic competition, and the possibility that a candidate may be evaluated within a complex botanical or compound-mixture context. Interaction prediction supports risk identification, but the predicted association does not establish clinical magnitude, direction, or relevance.

Herb–drug interaction assessment is especially important because heterogeneous terminology, multicomponent preparations, incomplete constituent information, and uneven evidence quality can complicate model training and interpretation. Dedicated computational approaches may help organise and predict herbal medicine–drug relations, but their outputs remain dependent on data provenance, entity resolution, context, and validation [14]. Every toxicity or interaction prediction should also be accompanied by uncertainty and applicability-domain information, because confidence varies with data density, endpoint definition, molecular novelty, and model architecture [15]. Out-of-domain or highly uncertain predictions should trigger abstention, redesign, or direct experimental planning rather than favourable default interpretation. The governing boundary is therefore explicit: computational screening can reduce unsafe prioritization and identify evidence gaps, but it cannot independently justify a safety claim.

Table 1 summarises toxicity boundary and interaction-risk requirements for safety-by-design generative chemistry in natural product-inspired drug discovery.

Table 1. Toxicity Boundaries and Interaction-Risk Requirements for Safety-by-Design Generative Chemistry in Natural Product-Inspired Drug Discovery: ADMET Prediction, Structural Alerts, Reactive Groups, Off-Target Risk, Safety Pharmacology, Herb–Drug Interaction Risk, Drug–Drug Interaction Risk, Metabolism-Related Risk, Bioactivation Concern, Exposure Relevance, Uncertainty, and Claim Boundaries

Safety component	Core safety question	Required evidence or input	Safety-by-design function	Potential output	Risk of overinterpretation	Validation requirement	Decision boundary
ADMET prediction	Does the generated structure	Standardised structure; endpoint-	Screens multiple developability and safety dimensions	Endpoint profile with confidence	Treating predicted properties as measured	Independent models, relevant external data,	Poor, uncertain, or out-of-domain

	present predicted absorption, distribution, metabolism, excretion, or toxicity liabilities?	specific models; training-domain information; relevant physicochemical descriptors	before prioritization	and domain status	pharmacokinetics or demonstrated safety	and experimental ADME or toxicology planning	profiles trigger redesign, deferral, or additional testing
Toxicity prediction	Is there computational evidence of endpoint-specific toxic potential?	Validated endpoint models; relevant assay labels; exposure context; uncertainty estimates	Flags candidates that may cross predefined research-stage toxicity boundaries	Endpoint-specific risk flags or probability ranges	Conflating a model score with a toxicological conclusion	Calibration assessment, external validation, mechanistic review, and experimental confirmation	No candidate receives a safety designation from prediction alone
Structural alert screening	Does the structure contain motifs associated with recognised toxicological liabilities?	Curated alert definitions; molecular structure; provenance and domain of each alert	Provides an interpretable warning layer before optimization or synthesis	Alert list with rationale and evidence class	Assuming every alert is causal or that absence of alerts indicates safety	Expert toxicology and medicinal-chemistry assessment; orthogonal models; experimental planning	Alerts initiate review but do not independently establish toxicity or safety
Reactive group screening	Are potentially electrophilic, unstable, or otherwise reactive functionalities present?	Structure-based rules; reactivity knowledge; metabolic context; anticipated exposure	Identifies possible covalent, instability, or bioactivation concerns	Reactive-group flag and review priority	Inferring biological harm without reaction, exposure, or metabolic evidence	Chemical reactivity testing and context-specific toxicology where justified	Unresolved high-concern reactivity prevents routine prioritization
Off-target risk	Could the molecule plausibly engage unintended proteins or pathways?	Predicted target profiles; chemical and biological similarity; counterscreen data; concentration context	Extends safety review beyond intended-target activity	Ranked off-target hypotheses with uncertainty	Treating similarity, docking, or prediction as demonstrated binding or toxicity	Orthogonal computational methods and targeted experimental counterscreens	High-consequence or poorly bounded off-target hypotheses require escalation
Safety pharmacology concern	Could intended or unintended activity affect critical physiological systems?	Target and pathway evidence; relevant assay data; predicted exposure; pharmacological context	Prioritizes cardiovascular, neurological, respiratory, or other system-level concerns	Safety-pharmacology concern map	Inferring organism-level effects directly from molecular predictions	Expert safety-pharmacology review and appropriate experimental studies	Unresolved critical-system concerns prevent advancement beyond hypothesis status
Herb–drug interaction risk	Could a natural product-inspired candidate or associated botanical context alter another	Herbal constituent identity; enzyme, transporter, target, and interaction data; clinical-context evidence	Separates natural product-related interaction assessment from intrinsic molecular toxicity	Interaction hypotheses with evidence provenance and uncertainty	Assuming a predicted interaction is clinically significant	Mechanistic assays, exposure analysis, and clinically relevant evaluation where appropriate	No clinical interaction claim is allowed from computational prediction alone

	therapy's effect or exposure?						
Drug–drug interaction risk	Could the candidate affect or be affected by co-administered drugs?	Enzyme and transporter profiles; target interactions; exposure estimates; known interaction networks	Identifies pharmacokinetic and pharmacodynamic interaction hypotheses	Pairwise risk flags and plausible mechanisms	Treating database associations as causal or quantitatively predictive	Dedicated interaction assays and exposure-informed evaluation	Material unresolved interaction risk triggers redesign or targeted validation
Metabolism-related risk	Could metabolic clearance, inhibition, induction, or instability create a safety liability?	Enzyme and transporter predictions; metabolite hypotheses; microsomal or cellular evidence; exposure context	Screens for metabolic liabilities and interaction mechanisms	Metabolism-risk profile and validation plan	Treating predicted metabolism as experimentally established	In vitro metabolism studies and metabolite identification where warranted	Unsupported or high-risk metabolism predictions cannot support a safety claim
Bioactivation concern	Could metabolism produce reactive or toxic species?	Structural alerts; metabolite predictions; reactivity evidence; enzyme context	Flags candidates requiring deeper metabolic-toxicology review	Bioactivation hypothesis and evidence gaps	Assuming that a predicted metabolite is formed or toxic in vivo	Metabolite characterization, reactivity assessment, and toxicological evaluation as appropriate	High-concern unresolved bioactivation blocks advancement
Exposure relevance	Are predicted hazards plausible at attainable biological exposure?	Potency estimates; pharmacokinetic evidence or predictions; protein binding; tissue distribution; concentration context	Connects hazard signals to plausible exposure without dismissing uncertainty	Exposure-qualified concern level	Equating low predicted exposure with absence of risk	Experimental pharmacokinetic and dose–response evidence	Hazard conclusions remain provisional until exposure is empirically characterized
Dose or concentration boundary where supported	At what exposure does a safety-relevant effect emerge or remain unobserved?	Reliable dose–response or concentration–response evidence; assay context; uncertainty intervals	Prevents binary safe-or-unsafe interpretations when quantitative evidence exists	Evidence-qualified testing or operating range	Inventing thresholds or extrapolating beyond the studied system	Replicated concentration–response evidence and appropriate translation studies	Boundaries apply only to the endpoint and system that generated them
Toxicophore avoidance	Can known high-concern structural liabilities be excluded during generation?	Curated toxicophore definitions; applicability context; chemical representation	Applies negative-design constraints during generation and optimization	Rejected structures or redesign prompts	Treating toxicophore avoidance as sufficient proof of safety	Post-generation screening, expert review, and experimental testing	Avoidance reduces prioritization of known liabilities but does not create a safe candidate
Uncertainty estimation	How reliable is each safety or interaction prediction?	Model ensembles or calibrated uncertainty	Qualifies safety-relevant outputs with confidence	Confidence interval, uncertainty	Assuming numerical uncertainty is	Calibration testing under relevant	High or unverified uncertainty prevents

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

		methods; data density; prediction context	or abstention information	score, or abstention flag	always calibrated or complete	distribution shifts	confident prioritization
Applicability domain	Is the generated compound represented sufficiently by the model's training experience?	Chemical-space similarity; feature coverage; leverage or distance measures; endpoint-specific training data	Prevents extrapolative predictions from being presented as reliable evidence	In-domain, borderline, or out-of-domain designation	Using one similarity threshold as universal proof of validity	Multiple domain diagnostics and prospective error monitoring	Out-of-domain outputs are deferred or escalated for direct experimental evaluation
No safety-claim boundary	Does the available evidence justify describing the candidate as safe?	Integrated predictions, uncertainty, provenance, experimental results, and expert review	Establishes a hard communication and decision boundary	Candidate remains a safety-screened hypothesis rather than a safe candidate	Converting favourable screening results into therapeutic or clinical claims	Comprehensive experimental toxicology and subsequent development evidence	The framework never independently establishes safety, clinical utility, or regulatory readiness

Validation gates and human oversight

Validation gates translate safety-by-design from a collection of predictive tools into a controlled decision process. Before a generated molecule advances, the relevant models should be evaluated for reproducibility, calibration, endpoint suitability, and performance under chemical-domain shift. Uncertainty estimates are useful only when their behaviour has been examined across relevant molecular-property tasks and when they can distinguish comparatively supported predictions from unstable or extrapolative outputs [16]. An in silico gate should therefore consider applicability-domain status, out-of-distribution indicators, ensemble disagreement, sensitivity to molecular representation, and concordance across orthogonal methods. A candidate with favourable predictions but high uncertainty or weak domain support should be deferred, redesigned, or directed toward carefully planned experimental investigation rather than advanced by default.

Data provenance and model documentation constitute an earlier gate that determines whether subsequent predictions are interpretable at all. Realistic assessment of artificial intelligence in drug discovery requires meaningful comparators, prospective relevance, and evidence standards aligned with the intended decision, while chemical and biological datasets must be examined for assay context, duplication, leakage, scaffold bias, class imbalance, confounding, and label quality [17, 18]. For natural product-inspired chemistry, provenance review should additionally verify source structures, stereochemical assignments, constituent identities, curation history, and the relationship between a generated scaffold and its claimed natural product basis. Model documentation should state intended use, excluded uses, training-data composition, endpoint definitions, version history, performance limitations, domain restrictions, and known failure conditions. These records allow reviewers to distinguish a technically reproducible prediction from one that is scientifically appropriate for a particular safety decision.

Explainability may strengthen expert review by showing which molecular features, learned representations, biological relationships, or comparator compounds influenced a prediction, but an explanation is not equivalent to a causal mechanism or experimental validation [19]. Expert medicinal chemistry review should examine chemical plausibility, stability, synthetic accessibility, stereochemical integrity, reactivity, aggregation potential, and optimisation artefacts. Expert toxicology review should evaluate structural alerts, predicted hazards, endpoint relevance, exposure assumptions, bioactivation concerns, uncertainty, and evidence gaps. Safety pharmacology and interaction-risk reviewers should separately consider critical physiological systems, off-target hypotheses, metabolic pathways, transporter effects, pharmacodynamic overlap, and potential herb–drug or drug–drug interactions. Their conclusions should be documented in a decision audit trail that records disagreements, requested analyses, conditional approvals, and the rationale for any human override.

Human oversight is most valuable when it is embedded within explicit authority and stopping rules rather than

added as a final endorsement step. Molecular-generation systems may exploit objective functions, produce chemically implausible solutions, or satisfy narrow metrics while violating broader design intentions, making failure-mode testing and human rejection authority essential [20]. Prospective workflows that connect generation with synthesis and experimental testing illustrate why computational proposals should pass through sequential evidence gates and why results should feed back into subsequent design cycles [21]. Even when a generated molecule is synthesised and a targeted bioactivity hypothesis is experimentally supported, that evidence remains specific to the tested context and does not establish comprehensive safety, efficacy, developability, or clinical utility [22]. Candidate prioritization should therefore authorise only the next research action, while human reviewers retain the ability to stop, defer, modify, or reverse any model-supported decision.

Proposed safety-by-design framework

The proposed framework begins with research-objective definition and natural product scaffold selection. The research objective should specify the intended target or phenotype, the reason for using natural product-inspired chemistry, the acceptable scope of molecular novelty, the safety questions that must be addressed, and the outcomes that would require rejection or redesign. Verified natural product identity, chemical representation, stereochemistry, source provenance, and scaffold rationale should then be documented before generation. Molecular-generation quality should not be evaluated through one objective alone; validity, uniqueness, novelty, distributional similarity, and goal-directed behaviour provide complementary computational perspectives on whether a generative system is producing meaningful candidates [23]. These evaluations form an initial in silico gate, but they do not establish biological relevance or safety.

The second stage links molecular generation with natural product-likeness assessment, synthetic-accessibility review, bioactivity hypothesis screening, and multi-objective optimisation. The generative engine may propose structures conditioned on a scaffold, target, phenotype, or selected molecular properties, but the optimisation space should contain explicit penalties, constraints, and abstention rules for known or predicted liabilities. Contemporary molecular-generation research indicates that architecture, representation, optimisation strategy, evaluation, synthesis feasibility, and experimental testing are interdependent rather than separable design problems [24]. The proposed framework therefore avoids a linear "generate and rank" workflow. Instead, every generated structure is accompanied by its design rationale, provenance, objective profile, uncertainty, and unresolved evidence requirements.

The third stage applies safety constraint filtering, toxicity-boundary assessment, interaction-risk assessment, and integrated ADMET and off-target review. Safety constraints may exclude recognised toxicophores, reactive groups, chemically unstable motifs, or candidates with strongly unsupported property profiles, while softer constraints may trigger redesign or additional evidence gathering rather than automatic rejection. Natural product-inspired discovery may operate in sparse-data settings in which limited target, assay, or scaffold-specific evidence restricts model generalisation [25]. For this reason, the framework requires narrow applicability claims, separate reporting of each endpoint, and explicit identification of candidates that are distant from relevant training data. Safety constraints should be revised as validated evidence accumulates, but evaluation datasets and decision records should remain protected from uncontrolled feedback contamination.

The fourth stage examines model reliability and sequences validation gates. Toxicity models should be assessed not only through discrimination or ranking performance but also through calibration, error control, and the conditions under which predictions can be considered sufficiently reliable for a defined research decision [26]. Uncertainty should be incorporated into the evidence record and used to support abstention, escalation, or direct testing when predictions are unsupported or potentially out of distribution [27]. The proposed gate sequence comprises data and provenance review, model documentation, applicability-domain and bias assessment, in silico validation, experimental planning, in vitro evaluation, and conditional consideration of in vivo studies where scientifically necessary and ethically justified. Experimental toxicology, safety pharmacology, metabolism, reactivity, and interaction studies should be selected according to the specific evidence gaps identified by earlier gates.

The final stage integrates multidisciplinary expert review, candidate prioritization boundaries, feedback updating, and claim restrictions. Human expertise remains responsible for defining objectives, interpreting model outputs, resolving conflicting evidence, approving validation plans, and exercising override authority [28]. Natural product-inspired generative chemistry can broaden molecular hypothesis generation, but its products remain design proposals requiring medicinal-chemistry, pharmacological, and safety evaluation [29]. A candidate may be prioritised for further research only when its evidence dossier, uncertainty, feasibility, unresolved liabilities, and validation plan have been reviewed transparently. Rejected or deferred candidates should

inform refined constraints, improved documentation, revised validation plans, or appropriately governed model updates. Throughout the process, a no-safety-claim boundary prevents favourable computational or early experimental evidence from being described as proof of safety, while a no-clinical-claim boundary prevents discovery-stage outputs from being presented as clinically actionable or regulator-ready.

Table 2 presents the proposed safety-by-design framework for generative chemistry applied to natural product-inspired discovery.

Table 2. Proposed Safety-by-Design Framework for Natural Product-Inspired Generative Chemistry: Research Objective, Scaffold Selection, Molecular Generation, Natural Product-Likeness, Bioactivity Hypothesis, Safety Constraints, Toxicity Boundaries, Interaction-Risk Assessment, ADMET Review, Validation Gates, Expert Oversight, Feedback, and Claim Boundaries

Framework stage	Core purpose	Required input	Safety-by-design function	Potential output	Validation need	Failure risk	Feedback mechanism	Decision boundary
Research objective definition	Define the intended scientific question, use context, and unacceptable outcomes	Target or phenotype rationale; natural product context; intended research use; safety priorities	Converts broad discovery goals into bounded objectives, exclusions, and evidence requirements	Versioned design brief with endpoint hierarchy	Multidisciplinary review of scientific and safety relevance	Optimisation of an ambiguous, incomplete, or clinically overstated objective	Revise objectives when evidence, feasibility, or risk assumptions change	Generation does not begin without an approved research-purpose definition
Natural product scaffold selection	Select traceable natural product identities or scaffold hypotheses	Verified structures; stereochemistry; source and provenance; known biology; chemical tractability	Prevents misidentified or poorly characterised source chemistry from anchoring generation	Approved scaffold set with provenance record	Identity verification and expert natural product chemistry review	Propagation of incorrect structures, annotations, or unsupported traditional-use assumptions	Correct source records and update scaffold eligibility	Natural origin is never treated as evidence of safety
Molecular generation	Produce candidate hypotheses within the approved objective space	Scaffold representation; generative model; permitted design space; constraint definitions	Generates diverse structures while retaining explicit restrictions	Versioned candidate set with design rationale	Reproducibility, validity, duplication, memorisation, and failure-mode checks	Invalid structures, uncontrolled novelty, memorisation, or objective exploitation	Retrain, recondition, or revise objectives using documented evidence	Generated structures remain hypotheses only
Natural product-likeness assessment	Determine whether candidates preserve the intended natural product-inspired rationale	Structural descriptors; reference distributions; scaffold relationships; stereochemical features	Distinguishes natural product-inspired design from unrestricted generation	Natural product-likeness profile with uncertainty	Multiple similarity measures and expert chemical review	Treating a likeness score as identity, activity, or safety evidence	Adjust generation constraints and reference sets	Likeness supports design rationale only
Bioactivity hypothesis screening	Assess whether candidates plausibly address the intended target or phenotype	Activity models; target relevance; assay evidence; uncertainty; applicability-domain status	Removes candidates lacking a defensible activity hypothesis	Bioactivity hypotheses with evidence limitations	External validation, orthogonal methods, and experimental assay planning	Overreliance on docking, similarity, or model scores	Incorporate assay evidence and revise models or objectives	Prediction does not establish biological activity

Safety constraint filtering	Apply predefined negative and positive safety-related constraints during or after generation	ADMET models; toxicity models; structural alerts; reactive-group rules; uncertainty	Reduces prioritization of candidates with recognised or predicted liabilities	Filtered candidate set with documented exclusion or escalation reasons	Constraint sensitivity analysis and expert review	False reassurance, excessive filtering, or metric exploitation	Refine constraints using validated evidence and monitor rejected classes	Passing filters does not establish safety
Toxicity-boundary assessment	Integrate endpoint-specific hazard signals with confidence and context	Toxicity predictions; alerts; assay data; exposure assumptions; applicability-domain evidence	Establishes provisional research-stage limits for escalation, redesign, or rejection	Toxicity evidence profile and unresolved-risk register	Calibration, orthogonal evidence, mechanistic review, and experimental toxicology	Binary safe-or-unsafe interpretation of incomplete evidence	Update boundaries when relevant experimental evidence becomes available	No global safety conclusion is permitted
Interaction-risk assessment	Evaluate herb–drug, drug–drug, metabolic, transporter, and pharmacodynamic interaction hypotheses	Interaction evidence; target and metabolism information; constituent identity; exposure context	Adds an independent layer not captured by intrinsic toxicity prediction	Interaction-risk map with validation priorities	Mechanistic and exposure-informed experimental assessment	Missing complex, mixture-related, or context-dependent interactions	Update interaction models and records with verified evidence	Predicted interactions cannot be presented as clinical outcomes
ADMET and off-target review	Examine developability, metabolism, exposure, unintended targets, and safety-pharmacology concerns	Multi-endpoint ADMET profile; off-target predictions; biological signatures; concentration context	Detects cross-domain liabilities and competing optimisation objectives	Integrated liability profile	Endpoint-specific validation and counterscreen planning	Collapsing conflicting endpoints into a misleading aggregate score	Reweight objectives and redesign candidates based on evidence	Critical unresolved liabilities prevent routine prioritization
Uncertainty and applicability-domain review	Determine whether predictions are sufficiently supported to inform a research decision	Calibrated uncertainty; domain diagnostics; data density; ensemble disagreement	Enables abstention and prevents unsupported extrapolation	In-domain, borderline, out-of-domain, or indeterminate status	Calibration and shift-specific performance assessment	Confident errors on novel natural product-inspired chemistry	Expand relevant data, retrain models, or prioritise direct testing	Out-of-domain candidates cannot pass solely on predicted scores
Validation gate sequencing	Organise computational, experimental, and expert reviews in a defensible order	Evidence from prior stages; predefined gate criteria; stopping rules	Prevents selective advancement and separates evidence levels	Advance, redesign, defer, reject, or test decision	Independent review of criteria, reproducibility, and relevance	Movement from generation directly to strong safety or efficacy claims	Use failed gates and new evidence to revise earlier stages	Each gate authorises only the next research action
Expert review	Integrate medicinal chemistry, toxicology, pharmacology, natural product, and data-science judgment	Full evidence dossier; explanations; uncertainties; unresolved conflicts	Detects contextual failures and weighs evidence that cannot be reduced to one score	Documented multidisciplinary recommendation	Transparent rationale, conflict review, and recording of dissent	Automation bias, disciplinary blind spots, or undocumented intuition	Record overrides, dissent, and requested analyses	Experts may stop, defer, or condition advancement

Candidate prioritization	Select candidates for further research under explicit evidence and uncertainty constraints	Integrated activity hypothesis; safety signals; feasibility; gate status; validation plan	Directs resources without implying safety or therapeutic utility	Prioritised research candidates with explicit conditions	Multidisciplinary confirmation of remaining evidence needs	Candidate rank interpreted as proof of quality, safety, or readiness	Reassess after new experimental or model evidence	Prioritization supports research planning only
Feedback updating	Learn from synthesis, assay, toxicology, interaction, and review outcomes	New experimental evidence; rejected candidates; reviewer decisions; domain-shift information	Improves models, constraints, documentation, and gate criteria while preserving version history	Updated models, constraints, records, or validation plans	Prospective monitoring and version-to-version comparison	Feedback bias, selective retention, or contamination of evaluation data	Controlled update cycle with frozen benchmarks and audit logs	Updated systems require revalidation before reuse
No safety-claim boundary	Prevent favourable screening evidence from becoming a safety assertion	Predictive and experimental evidence with limitations and uncertainty	Enforces disciplined language and decision scope	Safety-screened candidate hypothesis, where justified	Toxicology and governance review of wording and evidence	Misrepresentation of predictions as demonstrated safety	Correct communications and strengthen missing evidence	The framework does not independently establish safety
No clinical-claim boundary	Prevent discovery outputs from being presented as clinically actionable or regulator-ready	Candidate dossier; development stage; clinical evidence status	Separates discovery prioritization from therapeutic decision-making	Explicit nonclinical research-use limitation	Clinical, ethical, and regulatory assessment outside the framework	Premature clinical inference, use, or unsupported translation	Reassess only when appropriate development evidence exists	No claim of clinical safety, efficacy, utility, or regulatory readiness

Figure 1 illustrates the proposed safety-by-design framework for natural product-inspired generative chemistry, linking molecular generation, toxicity boundaries, interaction-risk assessment, validation gates, human oversight, and feedback updating.

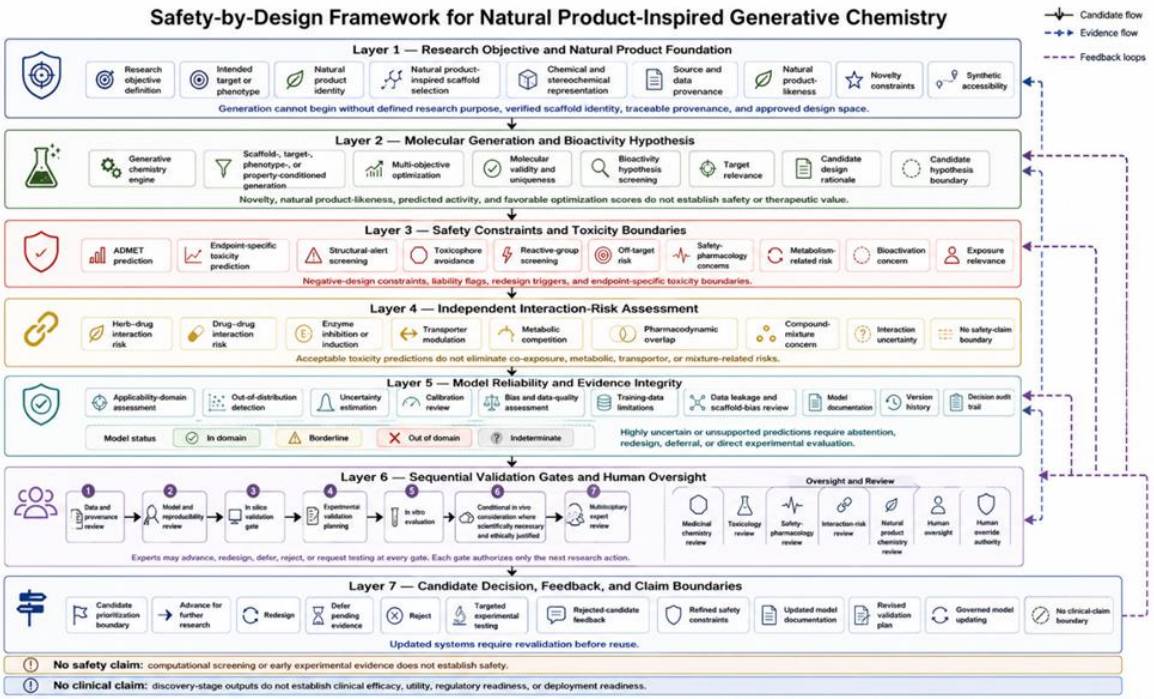

Figure 1. Safety-by-Design Framework for Natural Product-Inspired Generative Chemistry

Alt-text description

A conceptual safety-by-design framework showing natural product-inspired molecular generation flowing through safety constraints, toxicity-boundary assessment, interaction-risk review, ADMET and off-target evaluation, uncertainty assessment, validation gates, human oversight, feedback loops, and no safety-claim boundaries.

CONCLUSION

This article proposes a safety-by-design framework for natural product-inspired generative chemistry in which molecular generation is linked prospectively to toxicity boundaries, ADMET constraints, interaction-risk assessment, model-reliability checks, sequential validation gates, expert review, human override, and evidence feedback. The framework treats natural product identity, natural product-likeness, molecular novelty, predicted bioactivity, and synthetic accessibility as components of candidate-hypothesis formation rather than proof of safety or therapeutic value. It further separates intrinsic toxicity assessment from herb–drug, drug–drug, metabolic, off-target, and safety-pharmacology risks, while requiring uncertainty, applicability-domain status, provenance, and model limitations to remain visible throughout decision-making. Candidate prioritization is restricted to determining whether further research is justified, and explicit no-safety-claim and no-clinical-claim boundaries prevent computational or early experimental findings from being overstated. Generated molecules should therefore remain research hypotheses until their identity, mechanism, efficacy, exposure, interaction profile, toxicology, and developability have been evaluated through appropriate experimental and translational evidence.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5

[2] Sánchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018;361(6400):360-5. doi:10.1126/science.aat2663

[3] Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3

[4] Walters WP, Murcko M. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol. 2020;38(2):143-5. doi:10.1038/s41587-020-0418-2

[5] Meyers J, Fabian B, Brown N. De novo molecular design and generative models. Drug Discov Today. 2021;26(11):2707-15. doi:10.1016/j.drudis.2021.05.019

[6] Gao W, Coley CW. The synthesizability of molecules proposed by generative models. J Chem Inf Model. 2020;60(12):5714-23. doi:10.1021/acs.jcim.0c00174

[7] Merk D, Grisoni F, Friedrich L, Schneider G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. Commun Chem. 2018;1:68. doi:10.1038/s42004-018-0068-1

[8] Shen X, Zeng T, Chen N, Li J, Wu R. NIMO: A natural product-inspired molecular generative model based on conditional Transformer. Molecules. 2024;29(8):1867. doi:10.3390/molecules29081867

[9] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: A benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30. doi:10.1039/C7SC02664A

[10] Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, et al. ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res. 2021;49(W1):W5-14. doi:10.1093/nar/gkab255

[11] Yang H, Sun L, Li W, Liu G, Tang Y. In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front Chem. 2018;6:30. doi:10.3389/fchem.2018.00030

[12] Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, et al. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. Nat Biotechnol. 2020;38(9):1087-96. doi:10.1038/s41587-020-0502-7

[13] Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug-drug and drug-food interactions. Proc Natl Acad Sci U S A. 2018;115(18):E4304-11. doi:10.1073/pnas.1803294115

[14] Yuan S, Zhou Z, Jin X, Zhuo L, Li K. Enhancing herbal medicine-drug interaction prediction using large language models. IEEE J Biomed Health Inform.

2025;29(10):6971-81. doi:10.1109/JBHI.2025.3558667

[15] 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

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

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

[18] Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data used for AI in drug discovery. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037

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

[20] Renz P, Van Rompaey D, Wegner JK, Hochreiter S, Klambauer G. On failure modes in molecule generation and optimization. Drug Discov Today Technol. 2019;32-33:55-63. doi:10.1016/j.ddtec.2020.09.003

[21] Grisoni F, Huisman BJH, Button AL, Moret M, Atz K, Merk D, et al. Combining generative artificial intelligence and on-chip synthesis for de novo drug design. Sci Adv. 2021;7(24):eabg3338. doi:10.1126/sciadv.abg3338

[22] Merk D, Friedrich L, Grisoni F, Schneider G. De novo design of bioactive small molecules by artificial intelligence. Mol Inform. 2018;37(1-2):1700153. doi:10.1002/minf.201700153

[23] Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096-108. doi:10.1021/acs.jcim.8b00839

[24] Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF. Generative models for molecular discovery: Recent advances and challenges. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(5):e1608. doi:10.1002/wcms.1608

[25] Moret M, Friedrich L, Grisoni F, Merk D, Schneider G. Generative molecular design in low data regimes. Nat Mach Intell. 2020;2(3):171-80. doi:10.1038/s42256-020-0160-y

[26] Morger A, Svensson F, Arvidsson McShane S, Gauraha N, Norinder U, Spjuth O, et al. Assessing the calibration in toxicological in vitro models with conformal prediction. J Cheminform. 2021;13(1):35. doi:10.1186/s13321-021-00511-5

[27] Yu J, Wang D, Zheng M. Uncertainty quantification: Can we trust artificial intelligence in drug discovery? iScience. 2022;25(8):104814. doi:10.1016/j.isci.2022.104814

[28] Schneider G. Mind and machine in drug design. Nat Mach Intell. 2019;1(3):128-30. doi:10.1038/s42256-019-0030-7

[29] Schneider P, Altmann KH, Schneider G. Generating bioactive natural product-inspired molecules with machine intelligence. Chimia (Aarau). 2022;76(5):396-401. doi:10.2533/chimia.2022.396