Pharmacophore

ISSN-2229-5402

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

HAS GENERATIVE MOLECULAR DESIGN PRODUCED BETTER MEDICINES OR MERELY MORE PLAUSIBLE MOLECULES? A CRITICAL EVIDENCE REVIEW

Viviana Rojas[1]*, Juan Herrera[2], Silvia Monroy[3], Luis Gutierrez[4]

1. ***Department of Generative Molecular Design and Critical Review, Faculty of Pharmacy, University of São Paulo, São Paulo, Brazil.***
2. ***Department of Better Medicines vs. Plausible Molecules, Faculty of Pharmacy, National University of Rosario, Rosario, Argentina.***
3. ***Department of Evidence Assessment in Generative Design, Faculty of Pharmaceutical Sciences, University of Concepción, Concepción, Chile.***
4. ***Department of Translational Value of Generative Chemistry, Faculty of Pharmacy, National University of Asunción, Asunción, Paraguay.***

ARTICLE INFO

Received:
28 February 2025
Received in revised form:
01 June 2025
Accepted:
07 June 2025
Available online:
28 June 2025

Keywords: Generative molecular design, De novo drug design, Molecular generation, Synthetic feasibility, Medicinal chemistry, Experimental validation

ABSTRACT

Generative molecular design has rapidly expanded the capacity of computational systems to propose novel chemical structures, optimize predicted molecular properties, and explore chemical spaces that would be difficult to enumerate manually. Yet the pharmaceutical meaning of these capabilities remains unsettled. A structurally valid or computationally optimized molecule is not necessarily synthesizable, pharmacologically selective, developable, safe, or therapeutically superior. This critical review examines how evidence for generative molecular design should be interpreted across progressively demanding stages of pharmaceutical relevance. The selected literature was appraised according to the definition of model success, representation and data dependence, objective-function validity, synthetic feasibility, medicinal-chemistry suitability, experimental confirmation, developability evidence, comparator quality, reproducibility, and downstream progression. The synthesis shows that generative systems have convincingly demonstrated molecular representation, novelty generation, distribution learning, and optimization against specified computational objectives. Evidence is also emerging that synthesis-aware workflows can produce experimentally realizable and biologically active compounds in selected discovery programmes. However, much of the literature remains concentrated at the level of benchmark performance, predictor-mediated optimization, or isolated prospective examples. Evidence for consistent gains in selectivity, safety, pharmacokinetics, lead quality, candidate progression, or clinical benefit is substantially weaker. The article therefore proposes a graded evidentiary interpretation in which plausible molecular generation, predicted pharmaceutical utility, experimental confirmation, preclinical viability, and improved medicines are treated as distinct claim levels. This framework is not presented as a validated standard, but as a critical tool for preventing computational success from being mistaken for pharmaceutical improvement. Progress will depend on stronger comparators, prospective failure-inclusive evaluation, synthesis execution, multidimensional property testing, calibrated uncertainty, independent replication, and longitudinal reporting beyond initial compound generation.

To Cite This Article: Rojas V, Herrera J, Monroy S, Gutierrez L. Has Generative Molecular Design Produced Better Medicines or Merely More Plausible Molecules? A Critical Evidence Review. Pharmacophore. 2025;16(3):65-74. https://doi.org/10.51847/7SqaIiPJ4n

Introduction

AI-assisted molecular design has expanded the range and speed of computable design proposals, but established analyses caution that pharmaceutical impact depends on valid data, realistic problem selection, experimental integration, and decision relevance rather than model novelty alone [1–3]. This distinction is particularly important in small-molecule discovery, where a model can produce chemically formatted outputs long before those outputs satisfy the interacting requirements of synthesis, potency, selectivity, exposure, safety, formulation, and therapeutic differentiation. Generative systems can therefore succeed as algorithms while remaining uncertain as pharmaceutical technologies. The central scientific question is not whether a model

Corresponding Author: Viviana Rojas; Department of Generative Molecular Design and Critical Review, Faculty of Pharmacy, University of São Paulo, São Paulo, Brazil. E-mail: viviana.rojas@usp.br.

can generate molecules, but what level of evidence is required before those molecules can reasonably be described as better compounds, viable leads, development candidates, or improved medicines.

Generative molecular design is especially vulnerable to overinterpretation because chemical and biological data limitations coexist with evaluation conventions that reward valid, novel, or optimized structures without establishing therapeutic superiority [4, 5]. Validity generally indicates that a generated representation obeys specified syntactic or chemical rules. Novelty usually denotes dissimilarity from a training set or reference collection. Optimization success ordinarily means that a model has increased one or more selected scores, often supplied by a learned predictor, docking function, structural heuristic, or composite reward. Each achievement may be technically meaningful, yet none independently establishes that the proposed structure can be synthesized efficiently, retains activity in an experimental system, avoids important liabilities, or improves a development decision.

The unresolved gap is therefore evidentiary rather than merely computational. Much of the field is evaluated through measures that are convenient, repeatable, and compatible with public datasets, whereas pharmaceutical decisions require evidence that is context-specific, experimentally grounded, and sensitive to trade-offs. A generated molecule may appear favourable because it occupies a novel region of a learned representation, satisfies a property predictor, or receives a high docking score. Such results remain conditional on the training distribution, molecular encoding, scoring function, target model, applicability domain, and filtering rules used to define success. When these dependencies are obscured, benchmark performance may be interpreted as evidence of medicinal-chemistry value even though the benchmark does not test the relevant scientific claim.

This critical review evaluates whether generative molecular design has produced better medicines or has primarily improved the generation of plausible molecular proposals. It does not treat the alternatives as mutually exclusive. Instead, it separates demonstrated computational capability from plausible utility, synthesis-aware design, experimental confirmation, preclinical viability, and therapeutic improvement. The article critically examines how novelty, validity, and optimization are operationalized; why chemical plausibility differs from synthetic and medicinal-chemistry feasibility; what evidence supports gains in potency, selectivity, safety, and developability; and how prospective studies should be interpreted relative to computational benchmarks. Its principal conceptual contribution is a proposed maturity model for claims of pharmaceutical improvement. The model is intended to organize evidence and limit overstatement, not to function as a validated predictive, regulatory, or deployment-readiness instrument.

Review Questions, Evidence Boundaries, and Critical Appraisal Approach

The first review question asks what, precisely, has improved when a generative model is reported to perform well. Earlier assessments of automated de novo design emphasized that producing candidate structures is only one component of a wider medicinal-chemistry process and should not be equated with completing drug design [6]. Contemporary appraisal similarly indicates that AI utility must be evaluated against the requirements of a defined pharmaceutical problem rather than inferred from technical sophistication alone [7]. The review therefore distinguishes five progressively stronger claims: that a model generates valid molecular representations; that it generates chemically plausible or novel structures; that it improves predicted objective values; that it contributes to experimentally confirmed compounds; and that it improves downstream pharmaceutical outcomes. These claims are related, but they are not interchangeable, and evidence at one level does not automatically justify language belonging to another.

The evidence boundary includes peer-reviewed studies and critical analyses that directly inform molecular generation, benchmark construction, representation dependence, optimization, synthesizability, medicinal-chemistry assessment, property prediction, experimental validation, and translational interpretation. Molecular representations are treated as an appraisal domain because they determine which chemical distinctions a model can encode, which similarities it preserves, and which errors or biases may remain invisible [8]. A model trained on string representations, molecular graphs, fragments, three-dimensional environments, or target-conditioned structures may generate materially different outputs even when evaluated under nominally similar tasks. Consequently, model architecture is not interpreted independently of data provenance, molecular encoding, target definition, and the conditions under which generated structures are filtered or ranked.

The third review question concerns whether optimization objectives correspond to pharmaceutical value. Small-molecule discovery is intrinsically multi-objective because potency, selectivity, solubility, permeability, metabolic stability, exposure, safety, novelty, synthesis, and intellectual-property considerations may conflict rather than improve together [9]. Critical appraisal must therefore identify whether an objective is experimentally measured or computationally predicted, whether it represents a decision-critical endpoint or a convenient proxy, how uncertainty is handled, and whether optimization causes deterioration in unmeasured properties. A model's ability to maximize a reward is considered evidence of algorithmic control over that reward, not evidence that the underlying compound has improved unless the reward has been independently connected to the intended pharmaceutical outcome.

The appraisal framework evaluates evidence according to claim specificity, comparator adequacy, data suitability, representation dependence, predictor validity, synthesis assessment, prospective testing, failure reporting, external replication, and downstream progression. Existing classifications of automated chemical design show the value of separating levels of automation and experimental integration rather than treating automation as a binary status [10]. The present review extends that reasoning to evidence maturity: computational generation, synthesis prediction, laboratory realization, pharmacological testing, developability assessment, and therapeutic benefit are examined as distinct evidentiary states. The resulting synthesis is explicitly interpretive. It identifies what the selected evidence supports, where conclusions remain conditional, and what

cannot yet be concluded. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop review questions, evidence boundaries, and critical appraisal approach within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Review questions, evidence boundaries, and critical appraisal approach

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Molecular representation and validity	String-, graph-, fragment-, descriptor-, or structure-based generation with explicit validity checks	The system can produce machine-readable or chemically admissible molecular representations	Strong for technical generation capability	Validity rules, sanitization procedures, tautomer handling, charge states, and filtering may differ across studies	No direct evidence of synthesis, activity, or developability
Novelty and chemical-space coverage	Similarity analysis, training-set comparison, distributional metrics, scaffold analysis, or latent-space examination	Generated molecules differ from a defined reference collection or expand a represented chemical region	Moderate and definition-dependent	Novelty depends on the reference database, similarity measure, split design, and memorization controls	Relationship between structural novelty and useful medicinal innovation remains uncertain
Computational optimization	Reinforcement learning, latent-space search, docking, QSAR, property prediction, or composite scoring	Selected computational objectives improve during generation	Strong for optimization of encoded objectives	Reward misspecification, predictor error, applicability-domain failure, and objective gaming	Transfer from score improvement to measured pharmaceutical improvement
Synthetic and medicinal-chemistry feasibility	Synthetic-accessibility heuristics, retrosynthesis, reaction-constrained generation, chemist review, or route execution	Proposed molecules appear more likely to be synthesizable or chemically workable	Moderate; stronger when route-aware or experimentally confirmed	Planner coverage, reaction-data bias, reagent assumptions, route fragility, and omitted purification constraints	Prospective evidence including failed synthesis attempts and route practicality
Pharmacological performance	Biochemical, biophysical, structural, cellular, selectivity, or target-engagement testing	Generated compounds display activity in a defined experimental context	Strong for individual experimentally tested compounds; limited for field-wide generalization	Candidate-selection bias, assay dependence, narrow target coverage, and incomplete controls	Matched prospective comparison with alternative design strategies
Developability and safety	ADME, pharmacokinetic, off-target, toxicology, formulation, exposure, or in vivo assessment	Generated compounds possess properties compatible with further development	Limited and uneven across the evidence base	Prediction-to-experiment gaps, endpoint selection, incomplete liability panels, and model-system dependence	Integrated, longitudinal evidence from generation through candidate progression
Pharmaceutical improvement	Comparative lead-quality, development, clinical, or therapeutic evidence	Generative design produces better leads, candidates, or medicines than available alternatives	Not established as a general field-level conclusion	Weak counterfactuals, selective success reporting, limited independent replication, and absent downstream follow-up	Comparative programme-level and clinical evidence demonstrating added value

How Generative Models Define Novelty, Validity, and Optimization Success

The field's early operational definitions of success developed through continuous latent-space generation, reward-driven molecular optimization, and standardized tests of validity, novelty, distribution learning, and goal attainment [11–13]. These developments were essential because they converted an open-ended design problem into measurable computational tasks. Validity could be assessed by determining whether a representation decoded into an admissible molecular structure. Novelty could be measured relative to a training set or reference collection. Optimization could be evaluated by testing whether generated molecules achieved higher scores on predefined objectives. These constructs enabled reproducible comparison and accelerated methodological development, but they also created a narrow evidentiary vocabulary in which a model could be described as successful without demonstrating that its outputs were experimentally realizable or pharmaceutically useful.

Validity is the least demanding of these constructs. For string-based models, it may denote syntactically acceptable notation and chemically processable structures; for graph-based systems, it may involve permitted atoms, bonds, valence states, and connectivity; for three-dimensional generators, it may additionally involve geometric consistency. None of these definitions establishes that a molecule is stable, isolable, synthesizable under practical conditions, free from reactive liabilities, or compatible with a medicinal-chemistry programme. Distribution-oriented benchmarking expanded evaluation beyond simple validity by examining whether generated sets resemble or differ from selected molecular collections [14]. This provides useful evidence about model behaviour, diversity, and coverage, but the interpretation remains conditional on the dataset. A generator that closely reproduces a benchmark distribution may be learning chemically meaningful patterns, reproducing dataset biases, or both.

Novelty is similarly relational rather than intrinsic. A molecule may be novel relative to a training set while remaining familiar within a larger public or proprietary chemical universe. Conversely, a structure that appears highly dissimilar under a chosen

fingerprint may preserve familiar pharmacophoric, topological, or synthetic features. Novelty metrics are also sensitive to scaffold definitions, stereochemical treatment, tautomer normalization, molecular standardization, and the similarity threshold chosen by the investigator. Accordingly, novelty should be interpreted as a documented relationship between a generated structure and a specified comparison set. It should not be treated as direct evidence of mechanistic originality, patentability, synthetic accessibility, target relevance, or therapeutic differentiation. Configurable scoring environments make this dependence especially visible because model rankings can change when objectives, filters, aggregation rules, or benchmark settings are altered [15].

Optimization success is the most easily overextended construct because it appears to align directly with drug-design goals. In practice, a generative model normally optimizes a formal representation of a goal rather than the goal itself. A reinforcement-learning system may maximize a predicted activity value, docking score, similarity constraint, synthetic-accessibility estimate, or weighted combination of properties. The resulting improvement demonstrates that the generator can exploit information encoded in the scoring system. It does not establish that the score is calibrated, causally related to the intended biological outcome, robust outside its training domain, or resistant to adversarial molecular solutions. The critical interpretation proposed here is therefore that validity, novelty, and optimization constitute computational evidence classes, not pharmaceutical endpoints. They are necessary for evaluating model behaviour and may support candidate prioritization, but claims of better compounds require evidence that survives synthetic, medicinal-chemistry, experimental, and developability scrutiny.

Chemical Plausibility Versus Synthetic and Medicinal-Chemistry Feasibility

Chemical plausibility is often assigned through structural filters, drug-likeness rules, fragment frequencies, substructure alerts, or similarity to known molecular collections. These criteria can remove malformed or obviously undesirable structures, but they do not determine whether a proposed molecule can be prepared through a practical sequence of reactions. Comparative analysis of molecules produced by generative systems has shown that favourable generative metrics can coexist with substantial synthetic difficulty [16]. The distinction is fundamental: a molecule may look chemically reasonable to a model while requiring unavailable starting materials, unsupported transformations, unstable intermediates, difficult stereochemical control, or purification steps that make laboratory execution unattractive.

Retrosynthesis-informed scoring strengthens feasibility assessment by asking whether a reaction-planning system can identify a path from accessible precursors to a proposed product. The retrosynthetic accessibility score was developed as a rapid learned approximation to more computationally demanding retrosynthetic planning and therefore offers a more chemically grounded signal than simple complexity heuristics [17]. Nevertheless, such a score remains indirect. It reflects the reaction knowledge, search procedure, precursor catalogue, and success definition of the underlying planner. A favourable classification indicates compatibility with that computational environment; it does not establish acceptable yield, route robustness, impurity control, operational safety, cost, scale, or reproducibility in a medicinal-chemistry laboratory.

Route-level planning provides a richer evidence layer because it exposes proposed disconnections, reaction templates, starting materials, and alternative synthetic paths. AiZynthFinder demonstrates how automated retrosynthetic search can support rapid evaluation of candidate structures and assist comparison of possible routes [18]. Yet identifying a route is not the same as validating one. Reaction databases overrepresent successful and commonly reported chemistry, while failed reactions, unusual substrates, problematic work-ups, and tacit laboratory knowledge are incompletely captured. Medicinal chemists also judge whether a route is sufficiently modular for analogue synthesis, whether key positions remain accessible for optimization, and whether the chemistry can support repeated design–make–test cycles rather than a single showcase synthesis.

Fast learned surrogates such as RetroGNN can make retrosynthesis-aware screening more scalable by approximating slower planning calculations across large generated libraries [19]. Their usefulness is therefore operational, especially when unconstrained generation produces more candidates than can be examined individually. However, surrogate estimates inherit the assumptions and errors of the planning systems used to create their labels. The critical boundary is that synthetic feasibility forms a hierarchy: structural heuristics provide weak preliminary evidence; route prediction provides stronger computational evidence; expert review adds context-sensitive medicinal-chemistry judgment; and prospective execution supplies the most direct confirmation. A viable pharmaceutical proposal must also permit analogue generation, stereochemical control, liability removal, and property optimization. Mere synthesizability, although essential, is not equivalent to medicinal-chemistry feasibility.

Evidence for Potency, Selectivity, Safety, and Developability Gains

Pharmaceutical quality is multidimensional, and improvement in one property can degrade another. Multi-objective drug design therefore requires explicit treatment of conflicting goals, uncertainty, and trade-offs rather than the unqualified maximization of a single reward [20]. A generated structure may receive a favourable potency prediction while becoming less soluble, more lipophilic, metabolically unstable, synthetically cumbersome, or vulnerable to off-target interactions. Scalar reward functions can hide these conflicts by combining heterogeneous objectives through weights that may not reflect project priorities. Pareto-based or constraint-aware approaches offer more transparent alternatives, but they still depend on the validity and applicability of each underlying predictor.

Developability models can assist early prioritization by estimating absorption, distribution, metabolism, excretion, toxicity, and related molecular properties across large candidate collections. Large-scale in silico ADMET modelling demonstrates the breadth of endpoints that can be predicted and the potential value of such tools for filtering compounds before resource-

intensive testing [21]. These predictions are not direct evidence that generated compounds possess favourable developability. Endpoint datasets differ in assay design, chemical coverage, measurement quality, censoring, and biological context. Predictions may also be correlated because they derive from related molecular features, creating the appearance of convergent evidence without genuinely independent support. Model confidence must therefore be distinguished from calibrated uncertainty and experimental confirmation.

Holistic compound-quality approaches attempt to extend evaluation beyond potency and ligand efficiency by including pharmacokinetic and physicochemical considerations [22]. This direction is scientifically preferable to potency-only optimization because medicinal chemistry advances compounds through balanced profiles rather than isolated maxima. Even so, composite scores require normative choices about weighting, thresholds, and acceptable compromises. A property that is critical for an orally administered chronic therapy may be less decisive for another route, duration, or indication. Consequently, developability cannot be reduced to a universal score. Generative objectives should be tied to a clearly defined target product profile, assay context, chemical series, intended route, and development stage.

Prospective multitarget design provides evidence that generative chemistry can move beyond single-score optimization. Deep generative methods have been used to nominate compounds intended to affect defined target combinations, followed by synthesis and experimental testing [23]. Such studies strengthen the case that generative systems can contribute to pharmacological hypothesis formation and compound prioritization. They do not establish general gains in safety or developability, because successful activity in selected assays leaves unresolved broader selectivity, cellular context, exposure, metabolism, toxicity, formulation, and in vivo performance. The available evidence therefore supports a bounded conclusion: generated compounds can demonstrate measured potency and designed polypharmacology in particular programmes, while consistent field-wide improvement in integrated compound quality remains unproven.

Experimental Validation and Progression beyond Computational Benchmarks

Prospective validation changes the evidentiary status of generative molecular design because nominated structures are subjected to synthesis and experimental measurement rather than judged only through computational proxies. The reported identification of DDR1 kinase inhibitors demonstrated that a generative workflow could produce compounds that were synthesized and found to possess target activity [24]. This study is important as a proof of capability, but its interpretation must remain proportional. It confirms that generation can contribute to a successful target-focused design episode; it does not establish that the approach consistently outperforms expert medicinal chemistry, virtual screening, analogue enumeration, or other computational design strategies across diverse projects.

A more extensive prospective example combined deep interactome learning with compound synthesis, biochemical testing, biophysical characterization, selectivity assessment, and structural investigation of generated nuclear-receptor ligands [25]. The integration of orthogonal evidence is particularly valuable because agreement across different experimental modalities is more informative than reliance on one potency assay. Structural or biophysical confirmation can test whether the intended interaction is plausible, while selectivity studies expose liabilities that a target-specific score may miss. Even this stronger validation remains programme-specific. It does not alone demonstrate favourable pharmacokinetics, long-term safety, scalable synthesis, clinical differentiation, or routine generalizability to targets with different data availability and binding-site properties.

Generative design has also been connected to the design and validation of structurally novel, synthetically accessible antibiotic candidates [26]. Antibacterial discovery is a demanding context because molecular novelty, cell penetration, target engagement, spectrum, resistance, and toxicity can interact in ways that are poorly represented by simple molecular scores. Prospective synthesis and biological testing therefore provide meaningful evidence beyond novelty or predicted activity. However, an active generated compound remains an entry point for development rather than a completed medicine. Resistance potential, mammalian toxicity, exposure at relevant infection sites, formulation, dosing, and comparative efficacy must still be established before pharmaceutical improvement can be claimed.

Integration of generation with on-chip synthesis further reduces the separation between computational nomination and experimental realization by connecting design outputs to an executable chemical platform [27]. Such systems may increase the speed at which design hypotheses are tested and rejected, which could be more valuable than producing large untested virtual libraries. The strongest interpretation of the prospective evidence is therefore not that generative models have already transformed medicine development, but that they can participate productively in closed experimental cycles. Progress beyond benchmarks should be judged by the quality of nominated experiments, the inclusion of failures, the strength of comparators, the breadth of measured properties, and whether early activity survives subsequent lead optimization and development decisions.

A Maturity Model for Claims of Pharmaceutical Improvement

Realistic validation is difficult because public benchmarks, proprietary project data, medicinal-chemistry constraints, and prospective decision settings differ substantially. Comparative work using public and proprietary contexts shows that conclusions drawn from convenient benchmark tasks may not transfer reliably to project-realistic conditions [28]. The maturity model proposed here therefore classifies claims according to the strongest evidence actually demonstrated, rather than the sophistication of the model or the language used to describe it. It distinguishes representational validity, chemical plausibility, predicted optimization, synthesis-aware design, experimentally realized compounds, pharmacologically supported lead

hypotheses, preclinically viable candidates, and clinically demonstrated improvement. These levels are interpretive categories and have not been empirically validated as a universal standard.

Structure-based evaluation can raise the threshold above generic validity by asking whether generated molecules are compatible with a defined binding environment. Docking-oriented benchmarking illustrates that models producing valid and drug-like structures may still perform poorly when their outputs are evaluated against target-specific structural criteria [29]. Docking nevertheless remains a computational proxy affected by protein preparation, conformational sampling, scoring limitations, protonation choices, and the absence of explicit biological context. In the proposed model, docking or predicted affinity therefore belongs to the level of predicted objective improvement, not experimental pharmacological confirmation. **Figure 1** presents the generative molecule-to-medicine evidence funnel, showing how the article's key components and boundaries are connected within a maturity model for claims of pharmaceutical improvement.

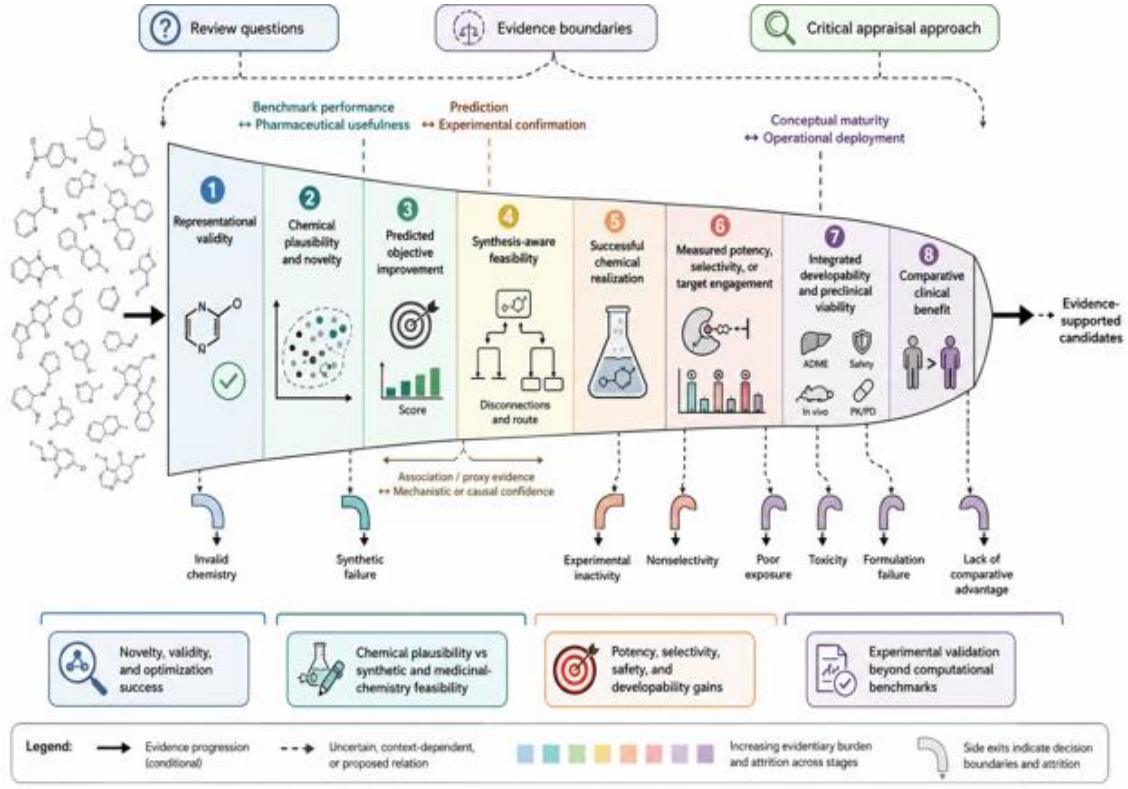

Figure 1. Generative Molecule-to-Medicine Evidence Funnel

The figure is an original conceptual synthesis that organizes the article's central contribution across review questions, evidence boundaries, and critical appraisal approach, review questions, evidence boundaries, critical appraisal approach, generative models define novelty, validity, and optimization success, generative models define novelty. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Optimization stress tests provide an additional boundary for the model. Curriculum and reinforcement-learning experiments show that success can depend strongly on task ordering, reward design, chemical representation, and the accessibility of the desired region of chemical space [30]. A generator that improves a benchmark objective may have learned a useful search strategy, exploited imperfections in a predictor, rediscovered familiar structures, or converged on a narrow class that satisfies the encoded score. The proposed maturity model therefore requires explicit reporting of the optimization target, predictor provenance, applicability domain, diversity constraints, failed runs, and comparison strategy before predicted improvement can support stronger claims.

At higher maturity levels, the relevant comparison changes from "Does the model generate attractive structures?" to "Does the workflow improve discovery decisions relative to credible alternatives?" Critical analysis of generative design and virtual screening suggests that once synthesizability and accessible chemical space are taken seriously, the distinction between generating a new molecule and selecting from very large make-on-demand spaces may become less clear [31]. A mature claim therefore requires a counterfactual: what would expert design, matched virtual screening, enumeration, or another optimization method have produced under the same constraints? Without such a comparator, experimental success demonstrates capability but not added value. The model consequently treats improved medicine as the strongest and least supported claim, requiring downstream comparative evidence rather than retrospective attribution to an AI component.

Research Priorities for Generation that Produces Viable Medicines

Future research should shift emphasis from producing ever larger virtual libraries toward generating decision-relevant hypotheses that can survive experimental and development constraints. Reviews of molecular generative modelling identify persistent challenges involving data quality, conditional generation, objective definition, model evaluation, synthesis, and experimental validation [32]. The most immediate priority is prospective, comparator-controlled evaluation. Candidate nomination criteria should be locked before testing, and generative workflows should be compared with credible alternatives using the same target information, compound budget, property constraints, and experimental resources. Unsuccessful syntheses, inactive compounds, and property failures should be reported alongside successes because selective presentation prevents realistic assessment of efficiency and risk.

Optimization research should make the relationship between learning strategy and pharmaceutical objective more transparent. Curriculum learning can expose a model to progressively more difficult constraints and may improve its ability to navigate complex optimization tasks [33]. Its pharmaceutical value, however, must be evaluated through external measurements rather than inferred from reward attainment. Research should therefore examine whether staged learning improves experimental hit quality, chemical diversity, route feasibility, uncertainty calibration, and robustness to changes in target or chemical series. Reward components should be traceable to defined decisions, and sensitivity analyses should show how candidate rankings change when predictors, weights, thresholds, or uncertainty penalties are altered.

Reproducibility also requires modular platforms that preserve the provenance of data, models, scoring components, filters, and generated candidates. REINVENT 4 illustrates the development of flexible generative infrastructure supporting several model and optimization strategies within a common environment [34]. Such platforms can facilitate controlled comparisons, but software availability alone does not guarantee reproducible pharmaceutical conclusions. Locked datasets, temporal splits, versioned predictors, standardized molecular preparation, independent reruns, candidate-level audit trails, and explicit reporting of human intervention are needed. Reproducibility should extend beyond regeneration of molecular files to replication of nomination decisions and experimental outcomes.

Three-dimensional and structure-conditioned generators represent an important frontier because they can incorporate binding-site geometry and molecular interactions more directly than unconditional generation. Equivariant diffusion methods demonstrate the computational feasibility of generating molecules in structural contexts and accommodating geometric constraints [35]. Their next test is not simply better structural benchmarks, but prospective comparison under protein flexibility, water networks, protonation uncertainty, induced fit, synthesis constraints, and assay-relevant biology. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop research priorities for generation that produces viable medicines within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Has Generative Molecular Design Produced Better Medicines or Merely More Plausible Molecules

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation
Unconditional molecular generation	Public molecular collections and distribution-learning tasks	Validity, uniqueness, novelty, diversity, and distribution similarity	Representational validity to chemical plausibility	Weak connection to a defined target, assay, synthesis route, or development need	Useful for model characterization and chemical-space exploration, but insufficient for claims of pharmaceutical improvement
Reward-driven property optimization	Predicted activity, docking, physicochemical, or composite objectives	Goal attainment, score improvement, constraint satisfaction, and diversity	Predicted objective improvement	Reward misspecification, predictor bias, applicability-domain failure, and objective gaming	Demonstrates control over encoded objectives, not necessarily improvement in measured properties
Multi-objective generation	Conflicting potency, selectivity, ADME, novelty, and synthesis goals	Pareto analysis, scalar rewards, constraint satisfaction, or rank aggregation	Predicted design advantage	Objective weighting may obscure trade-offs and project-specific priorities	Should be tied to a target product profile and experimentally measured endpoints
Synthesis-aware generation	Reaction data, retrosynthetic planners, precursor catalogues, or reaction-constrained spaces	Accessibility heuristics, route prediction, planner success, or chemist review	Synthesis-aware proposal	Route existence may not reflect yield, robustness, purification, scale, cost, or analogue flexibility	Stronger than chemical plausibility but subordinate to attempted synthesis
Closed-loop design–make–test systems	Defined targets with rapid synthesis and assay cycles	Prospective nomination, synthesis, testing, and iterative updating	Experimentally realized compound to pharmacologically supported lead hypothesis	Small programme scope, selective reporting, automation bias, and weak counterfactuals	Among the strongest demonstrations of practical capability when failures and comparators are included
Structure-conditioned and three-dimensional generation	Protein structures, pockets, complexes, or geometric constraints	Docking, pose quality, interaction recovery, structural validity, and prospective assays	Predicted target compatibility	Protein flexibility, scoring uncertainty, water, protonation, and synthetic feasibility	Promising for hypothesis generation but requires orthogonal experimental confirmation

Developability-aware generation	ADME, pharmacokinetic, safety, formulation, and exposure data	Endpoint prediction, uncertainty estimation, experimental panels, and longitudinal tracking	Predicted to preclinical viability	Sparse and heterogeneous data, correlated prediction errors, and context dependence	Integrated experimental profiles are required before developability gains can be claimed
Comparator-controlled prospective studies	Matched pharmaceutical projects and fixed resource budgets	Comparison with experts, virtual screening, enumeration, or alternative optimization methods	Added-value evidence	Difficult blinding, programme heterogeneity, and limited replication	Necessary to determine whether generation improves decisions rather than merely producing successful examples
Independent and multi-organization replication	Hidden targets, temporal test sets, proprietary data, and standardized protocols	Locked analysis, external validation, transfer assessment, and reproducibility audit	Generalizable capability evidence	Data-sharing constraints and variation in chemistry and assays	Required before routine pharmaceutical readiness can be inferred
Longitudinal programme follow-up	Progression from generated structure through optimization, candidate selection, and development	Attrition tracking, reason-for-failure reporting, decision impact, and comparative outcomes	Preclinical candidate to improved medicine	Long timelines, confounding by human and organizational contributions, and selective disclosure	The decisive evidence for whether generative design produces viable medicines rather than isolated plausible molecules

Conclusion

Generative molecular design has progressed beyond the production of arbitrary chemical strings: contemporary systems can generate valid structures, explore defined chemical spaces, optimize encoded objectives, incorporate synthesis and structural constraints, and, in selected programmes, contribute to compounds that are synthesized and experimentally active. These achievements justify serious scientific interest but not a field-wide claim that better medicines have already been produced. The strongest evidence remains concentrated at the levels of molecular plausibility, predicted optimization, and local experimental confirmation, whereas comparative lead quality, integrated developability, downstream candidate progression, and clinical superiority remain insufficiently established. The original contribution of this review is a proposed evidentiary maturity model that prevents representational validity, novelty, synthesis prediction, experimental activity, preclinical viability, and therapeutic improvement from being described as equivalent outcomes. Its boundaries are equally important: the model is interpretive rather than validated, does not assign universal development stages, and cannot substitute for project-specific scientific judgment. Generative design will become pharmaceutically consequential when it is evaluated through prospective counterfactual comparisons, failure-inclusive reporting, executable chemistry, orthogonal pharmacology, calibrated uncertainty, integrated developability testing, independent replication, and longitudinal evidence showing that early

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. 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
2. 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
3. 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
4. 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. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
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. Schneider G, Clark DE. Automated de novo drug design: Are we nearly there yet? Angew Chem Int Ed Engl. 2019;58(32):10792-803. doi:10.1002/anie.201814681
7. Hasselgren C, Oprea TI. Artificial intelligence for drug discovery: Are we there yet? Annu Rev Pharmacol Toxicol. 2024;64:527-50. doi:10.1146/annurev-pharmtox-040323-040828

8. David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: A review and practical guide. J Cheminform. 2020;12(1):56. doi:10.1186/s13321-020-00460-5
9. Fromer JC, Coley CW. Computer-aided multi-objective optimization in small molecule discovery. Patterns (N Y). 2023;4(2):100678. doi:10.1016/j.patter.2023.100678
10. Goldman B, Kearnes S, Kramer T, Riley P, Walters WP. Defining levels of automated chemical design. J Med Chem. 2022;65(10):7073-87. doi:10.1021/acs.jmedchem.2c00334
11. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76. doi:10.1021/acscentsci.7b00572
12. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminform. 2017;9(1):48. doi:10.1186/s13321-017-0235-x
13. 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
14. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular Sets (MOSES): A benchmarking platform for molecular generation models. Front Pharmacol. 2020;11:565644. doi:10.3389/fphar.2020.565644
15. Thomas M, O'Boyle NM, Bender A, de Graaf C. MolScore: A scoring, evaluation and benchmarking framework for generative models in de novo drug design. J Cheminform. 2024;16(1):64. doi:10.1186/s13321-024-00861-w
16. 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
17. Thakkar A, Chadimová V, Bjerrum EJ, Engkvist O, Reymond JL. Retrosynthetic accessibility score (RAscore)—rapid machine learned synthesizability classification from AI driven retrosynthetic planning. Chem Sci. 2021;12(9):3339-49. doi:10.1039/D0SC05401A
18. 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
19. Liu CH, Korablyov M, Jastrzębski S, Włodarczyk-Pruszyński P, Bengio Y, Segler MHS. RetroGNN: Fast estimation of synthesizability for virtual screening and de novo design by learning from slow retrosynthesis software. J Chem Inf Model. 2022;62(10):2293-300. doi:10.1021/acs.jcim.1c01476
20. Luukkonen S, van den Maagdenberg HW, Emmerich MTM, van Westen GJP. Artificial intelligence in multi-objective drug design. Curr Opin Struct Biol. 2023;79:102537. doi:10.1016/j.sbi.2023.102537
21. Beckers M, Sturm N, Sirockin F, Fechner N, Stiefl N. Prediction of small-molecule developability using large-scale in silico ADMET models. J Med Chem. 2023;66(20):14047-60. doi:10.1021/acs.jmedchem.3c01083
22. Tautermann CS, Borghardt JM, Pfau R, Zentgraf M, Weskamp N, Sauer A. Towards holistic compound quality scores: Extending ligand efficiency indices with compound pharmacokinetic characteristics. Drug Discov Today. 2023;28(11):103758. doi:10.1016/j.drudis.2023.103758
23. Munson BP, Chen M, Bogosian A, Kreisberg JF, Licon K, Abagyan R, et al. De novo generation of multi-target compounds using deep generative chemistry. Nat Commun. 2024;15(1):3636. doi:10.1038/s41467-024-47120-y
24. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40. doi:10.1038/s41587-019-0224-x
25. Atz K, Cotos L, Isert C, Håkansson M, Focht D, Hilleke M, et al. Prospective de novo drug design with deep interactome learning. Nat Commun. 2024;15(1):3408. doi:10.1038/s41467-024-47613-w
26. Swanson K, Liu G, Catacutan DB, Arnold A, Zou J, Stokes JM. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. Nat Mach Intell. 2024;6(3):338-53. doi:10.1038/s42256-024-00809-7
27. 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). doi:10.1126/sciadv.abg3338
28. Handa K, Thomas MC, Kageyama M, Iijima T, Bender A. On the difficulty of validating molecular generative models realistically: A case study on public and proprietary data. J Cheminform. 2023;15(1):112. doi:10.1186/s13321-023-00781-1
29. Ciepliński T, Danel T, Podlewska S, Jastrzębski S. Generative models should at least be able to design molecules that dock well: A new benchmark. J Chem Inf Model. 2023;63(11):3238-47. doi:10.1021/acs.jcim.2c01355
30. Mokaya M, Imrie F, van Hoorn WP, Kalisz A, Bradley AR, Deane CM. Testing the limits of SMILES-based de novo molecular generation with curriculum and deep reinforcement learning. Nat Mach Intell. 2023;5(4):386-94. doi:10.1038/s42256-023-00636-2
31. Stanley M, Segler M. Fake it until you make it? Generative de novo design and virtual screening of synthesizable molecules. Curr Opin Struct Biol. 2023;82:102658. doi:10.1016/j.sbi.2023.102658
32. Bilodeau CL, Jin W, Jaakkola TS, Barzilay R, Jensen KF. Generative models for molecular discovery: Recent advances and challenges. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(5). doi:10.1002/wcms.1608
33. Guo J, Fialková V, Arango JD, Margreitter C, Janet JP, Papadopoulos K, et al. Improving de novo molecular design with curriculum learning. Nat Mach Intell. 2022;4(6):555-63. doi:10.1038/s42256-022-00494-4

34. Loeffler HH, He J, Tibo A, Janet JP, Voronov A, Mervin LH, et al. Reinvent 4: Modern AI-driven generative molecule design. J Cheminform. 2024;16(1):20. doi:10.1186/s13321-024-00812-5
35. Schneuing A, Harris C, Du Y, Didi K, Jamasb A, Igashov I, et al. Structure-based drug design with equivariant diffusion models. Nat Comput Sci. 2024;4(12):899-909. doi:10.1038/s43588-024-00737-x