Pharmacophore

ISSN-2229-5402

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

GENERATIVE AI FOR DE NOVO DRUG DESIGN: A SYSTEMATIC REVIEW

Sophie Laurent[1]*, Pierre Dubois[1], Marc Lefevre[2], Claire Moreau[1]

1. ***Department of Computational Pharmaceutical Sciences, Faculty of Pharmacy, Sorbonne University, Paris, France.***
2. ***Department of Intelligent Drug Systems, Faculty of Medicine, École Polytechnique, Paris, France.***

ARTICLE INFO

Received:
14 February 2026
Received in revised form:
26 April 2026
Accepted:
28 April 2026
Available online:
28 June 2026

Keywords: Generative AI, De novo drug design, Molecular generation, Synthetic feasibility, Pharmacological validation, PRISMA 2020

ABSTRACT

Generative AI has become a prominent approach in de novo drug design because it can propose new chemical structures rather than only screen existing libraries. Across the past decade, the field has expanded from early autoencoder and reinforcement learning systems to graph, transformer, and diffusion-based molecular generators. This systematic review evaluated generative AI models applied to de novo molecular design. The review focused on model families, molecular representations, synthetic feasibility, pharmacological validation, evaluation metrics, and translational readiness. A systematic review was conducted according to PRISMA 2020 principles using searches of PubMed, Scopus, IEEE Xplore, and Web of Science. Records were screened by two reviewers, eligible studies were extracted using a structured form, and findings were synthesized narratively. The evidence base showed rapid methodological growth and frequent reporting of chemically valid, novel, and diverse molecules. However, synthetic feasibility was inconsistently integrated, and prospective experimental pharmacological validation was reported in only a small subset of studies. Generative AI for de novo drug design is technically sophisticated but remains incompletely translated into experimentally validated lead discovery. The main barriers are weak synthesis-aware optimization, inconsistent benchmarking, limited external validation, and scarce biological testing.

To Cite This Article: Laurent S, Dubois P, Lefevre M, Moreau C. Generative AI for De Novo Drug Design: A Systematic Review. Pharmacophore. 2026;17(3):1-11. https://doi.org/10.51847/em7LHgbmpo

Introduction

The drug discovery process remains long, costly, and uncertain, which has encouraged sustained interest in computational strategies that can improve early molecular design. Generative AI has been proposed as a way to move beyond virtual screening by creating novel chemical matter optimized toward desired biological and physicochemical properties [1, 2]. Several reviews have described the emergence of deep generative molecular design as a distinct area within computational drug discovery [3, 4], while broader analyses have emphasized that translation from algorithmic novelty to experimentally useful compounds remains limited [5]. In this context, systematic synthesis of the evidence is needed because the literature has expanded quickly across different model families, datasets, and validation practices.

The field has progressed from early variational autoencoders and recurrent sequence models toward generative adversarial networks, reinforcement learning, graph generators, transformers, and diffusion models. Early studies demonstrated that continuous latent representations could support molecular optimization [6], while reinforcement learning studies introduced property-directed molecular generation as a central paradigm [7, 8]. GAN-based and adversarial approaches later broadened the range of generative strategies but also introduced concerns about training stability and chemical realism [9, 10]. More recent transformer and diffusion architectures have further diversified the evidence base, particularly through language-based molecular generation and emerging three-dimensional generative approaches [11, 12].

A recurring challenge is that generated molecules must be more than valid strings or graphs; they must also be synthetically accessible, pharmacologically relevant, and suitable for downstream optimization. Benchmarking studies have standardized parts of this evaluation through metrics such as validity, uniqueness, novelty, diversity, and distributional similarity [13-16], but these criteria do not necessarily establish developability. Synthetic complexity and retrosynthetic accessibility methods

Corresponding Author: Sophie Laurent; Department of Computational Pharmaceutical Sciences, Faculty of Pharmacy, Sorbonne University, Paris, France. E-mail: sophie.laurent@gmail.com.

have been proposed to address this gap [17], yet several generative workflows continue to treat feasibility as a post-hoc filter rather than a design constraint. Experimental validation remains especially uncommon, although a small number of reports have shown that AI-generated compounds can progress to biochemical testing.

The objective of this systematic review was to evaluate generative AI models for de novo drug design from 2017 to 2026, with emphasis on architecture, molecular representation, synthetic feasibility, pharmacological validation, and translational readiness. The review was designed to integrate evidence from original research while contextualizing methodological trends identified in major reviews of molecular generation [1-5, 18-20]. It specifically examined whether reported advances in model sophistication were matched by improvements in synthesis-aware optimization and experimental testing. The review also followed PRISMA 2020 principles to structure search, screening, eligibility assessment, and narrative synthesis [1].

Materials and Methods

Search Strategy

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement. Searches were conducted in PubMed, Scopus, IEEE Xplore, and Web of Science for publications from January 1, 2017, to December 31, 2026, using terms that combined generative AI, de novo drug design, molecular generation, synthetic feasibility, reinforcement learning, diffusion models, transformers, and pharmacological validation. Search strings were informed by terminology used in major reviews of deep generative molecular design [1-5, 18, 19] and by benchmark studies that defined common molecular generation evaluation tasks [13-15]. The search targeted peer-reviewed literature in computational chemistry, medicinal chemistry, cheminformatics, machine learning, and drug discovery.

Inclusion and Exclusion Criteria

Studies were eligible if they reported an original generative model applied to the design or optimization of novel small molecules for drug discovery or if they provided directly relevant methodological evidence on molecular representation, synthesis scoring, or evaluation benchmarks. Eligible studies included VAEs, autoencoders, GANs, reinforcement learning systems, graph generators, transformer-based models, diffusion models, and hybrid architectures applied to molecular generation [6-10, 12, 21-28]. Studies were excluded if they used only discriminative prediction models, focused exclusively on protein generation, lacked a de novo molecular design component, were not peer reviewed, or were not available in English. Reviews and opinion articles were used for contextual framing where appropriate [1-5, 18-20], but the core synthesis prioritized primary generative-model studies and benchmark papers.

Screening and Selection

After database searching, 2,214 records were identified, and 1,842 remained after removal of duplicates and clearly irrelevant records. Title and abstract screening excluded 1,532 records, leaving 310 full-text articles assessed for eligibility, of which 105 were included in the qualitative synthesis. Common exclusion reasons at full text included absence of a generative component, use of only predictive QSAR or docking models, lack of drug-like molecular outputs, insufficient methodological detail, or non-peer-reviewed status. The PRISMA flow diagram for this process is reported as **Figure 1**, and the screening categories were aligned with prior benchmark and review conventions in molecular generation [13-16].

Figure 1 presents the PRISMA 2020 study-selection flow from 2,214 identified records to 105 studies included in the qualitative synthesis.

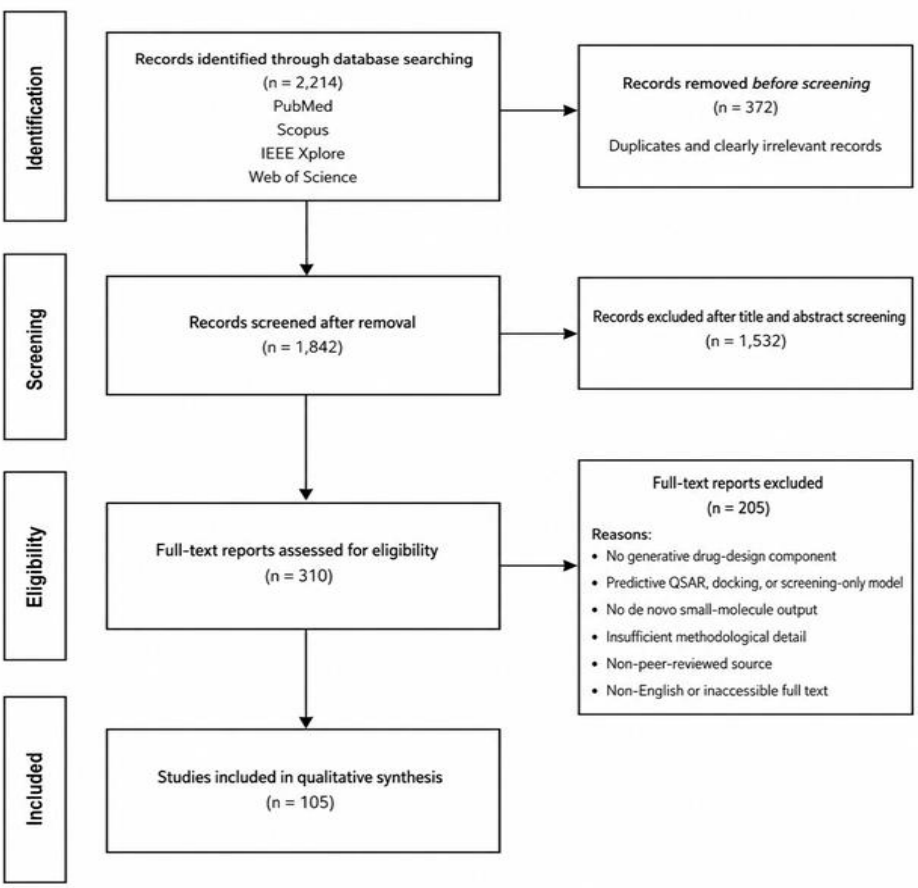

Figure 1. PRISMA 2020 Flow Diagram for Study Selection in the Systematic Review of Generative AI for De Novo Drug Design

Data Extraction

Data extraction captured publication year, model family, molecular representation, training dataset, optimization strategy, evaluation metrics, synthetic feasibility assessment, and validation level. Representation categories included SMILES and randomized SMILES [27], SELFIES [26], molecular graphs [24, 25], fragments, and three-dimensional molecular forms described in emerging diffusion-model work [11]. Synthetic feasibility variables included the use of synthetic accessibility scores, learned synthetic complexity, retrosynthesis-based feasibility, and whether such scores were integrated during generation or applied after generation. Pharmacological validation was extracted as in silico only, biochemical or cell-based testing, ADMET profiling, in vivo testing, or translational follow-up.

Risk of Bias Assessment

Risk of bias was assessed qualitatively because the included studies were heterogeneous and rarely designed as comparative intervention studies. Criteria addressed data leakage between training and evaluation sets, over-optimization to benchmark objectives, lack of external test sets, absence of prospective validation, inadequate reporting of failed generations, and failure to evaluate synthesis feasibility [13-16]. Studies that reported only internal validity, novelty, and property metrics without external biological or synthetic validation were judged to have limited translational evidence, even when algorithmic reporting was complete [14, 15]. Particular attention was given to whether feasibility scores such as SCScore or retrosynthetic accessibility were embedded in the generative loop rather than added only after molecule generation [17].

Synthesis Methods

A narrative synthesis was performed because differences in model architectures, datasets, targets, molecular representations, and validation endpoints precluded meta-analysis. Studies were grouped by generative model family, including VAEs and autoencoders [6, 24, 25], GANs and adversarial systems [9, 10, 23], reinforcement learning approaches [7, 8, 21, 22], diffusion models [11], and transformer or language-based systems [12, 28]. Vote-counting categories were used to summarize whether studies reported synthetic feasibility assessment, multi-objective optimization, standardized benchmark evaluation, or experimental pharmacological validation. The synthesis emphasized patterns across studies rather than pooled performance estimates, consistent with the methodological diversity described in prior reviews [1-5, 18-20].

Results and Discussion

Study Selection

The final synthesis included 105 studies from an initial set of 2,214 records identified across databases. After duplicate removal, title and abstract screening, and full-text review, most exclusions reflected lack of a generative drug-design component or exclusive reliance on predictive models rather than molecular generation. Studies retained for synthesis included early latent-space molecular design [6], reinforcement learning frameworks [7, 8], GAN-based approaches [9, 10], benchmark platforms [14, 15], and recent transformer or diffusion-based systems [11, 12, 28]. **Figure 1** should present the PRISMA flow with 2,214 records identified, 1,842 screened after deduplication, 310 full texts assessed, and 105 studies included.

Study Characteristics

Included studies were concentrated in the period after 2018, reflecting the rapid uptake of deep generative modeling in cheminformatics and drug discovery. Early work often used SMILES representations and latent-variable models [6, 27], while later studies increasingly used molecular graphs, SELFIES, transformers, and emerging three-dimensional representations [12, 24, 26]. Several papers were methodological demonstrations using public datasets and benchmark tasks [13-16], whereas a smaller subset included target-specific optimization or pharmacological follow-up [8]. Across the evidence base, most studies reported computational performance indicators, while fewer reported synthetic accessibility or experimental validation.

Model Types: VAEs and Autoencoders

Variational autoencoders and related autoencoder architectures formed an early foundation for de novo molecular design. The continuous latent representation introduced by Gómez-Bombarelli and colleagues enabled molecular interpolation and property-guided search in chemical space [6], and later graph-based autoencoder work adapted this paradigm to molecular graph generation [24, 25]. Several studies reported that these models could generate chemically plausible molecules, but reconstruction quality, latent-space smoothness, and validity depended strongly on representation and training data. In the reviewed literature, autoencoder models were often used to demonstrate feasibility of molecular generation rather than to establish experimentally validated drug discovery outcomes.

Model Types: Generative Adversarial Networks

Generative adversarial networks and adversarial autoencoders were applied to molecular design as a means of learning chemical distributions and generating molecules with desired properties. The druGAN model illustrated how adversarial autoencoding could be used to generate molecules with targeted physicochemical characteristics [9], while reinforced adversarial neural systems combined adversarial learning with property optimization [10]. Mol-CycleGAN extended adversarial concepts to molecular optimization by learning transformations between molecular domains [23]. Across the reviewed studies, GAN-based approaches were frequently described as innovative but were also associated with concerns about training instability, mode collapse, and uneven evaluation of synthetic feasibility.

Model Types: Reinforcement Learning

Reinforcement learning became a central strategy for steering molecular generators toward desired objectives after initial distribution learning. Olivecrona and colleagues reported a policy-based approach for molecular de novo design [7], and Popova and colleagues extended deep reinforcement learning toward goal-directed drug design [8]. REINVENT and its later versions provided influential open frameworks for property-directed molecular optimization and iterative design [21, 22]. In the reviewed evidence, reinforcement learning was especially prominent for multi-objective optimization, although reward design often risked over-optimizing proxy metrics rather than producing experimentally validated compounds.

Model Types: Diffusion Models

Diffusion models represented a newer and rapidly emerging model family in the generative drug-design literature. Recent reviews and methodological studies described diffusion approaches as promising for molecular generation because they can model complex distributions and may be adapted to three-dimensional molecular structures [11]. Compared with earlier SMILES-based models, diffusion approaches were more often discussed in relation to geometry-aware design and structure-conditioned generation. However, the reviewed evidence suggested that adoption in drug discovery remained earlier-stage than reinforcement learning or transformer-based approaches, with limited prospective validation.

Model Types: Transformer and Language-Based Models

Transformer and language-based architectures expanded molecular generation by treating chemical strings as sequences that can be modeled using methods related to natural language processing. MolGPT applied a transformer-decoder architecture to molecular generation and demonstrated the feasibility of autoregressive molecular language modeling [12]. Conditional transformer systems such as cMolGPT further linked generation to target-specific or property-conditioned design objectives [28]. In the reviewed literature, transformer models benefited from scalable sequence modeling but remained dependent on molecular representation choices, dataset quality, and the biological relevance of conditioning objectives.

Table 1 compares the major generative AI model families according to their design logic, molecular representation, translational evidence strength, and best-fit role within de novo drug discovery workflows.

Table 1. Translational Readiness Matrix across Generative AI Model Families in De Novo Drug Design

Generative model family	Primary design logic	Typical molecular representation	Main contribution to de novo design	Common evidence strength in reviewed literature	Key translational weakness	Best-fit role in drug discovery workflow
Variational autoencoders and autoencoders	Learn a continuous latent chemical space that can be searched or optimized	SMILES, molecular graphs, latent embeddings	Established early proof that molecular structures could be encoded, interpolated, and optimized computationally	Moderate evidence for molecular generation, validity, and latent-space optimization	Limited prospective biological validation; performance sensitive to representation quality and latent-space smoothness	Early-stage scaffold exploration and property-guided chemical-space navigation
Generative adversarial networks and adversarial autoencoders	Learn molecular distributions through generator–discriminator competition or adversarial regularization	SMILES, latent vectors, graph-derived features	Expanded generative diversity and molecular transformation approaches	Moderate methodological evidence, but uneven reproducibility across studies	Training instability, mode collapse, and inconsistent synthetic feasibility assessment	Distribution learning, chemical-space expansion, and exploratory molecular transformation
Reinforcement learning generators	Optimize molecule generation through reward functions linked to desired objectives	SMILES, graph actions, policy-based molecular construction	Became a dominant method for goal-directed and multi-objective molecular optimization	Strong in silico evidence for property optimization and benchmark performance	Vulnerable to reward hacking, proxy over-optimization, and weak experimental grounding	Hit-to-lead optimization when rewards include synthesis, ADMET, and assay-informed constraints
Graph-based generative models	Generate or modify molecular graphs using atom–bond structure directly	Molecular graphs, junction trees, graph neural representations	Improved structural chemical realism compared with purely string-based methods	Strong evidence for chemically structured generation and representation-aware modeling	May still lack route feasibility, assay validation, and target-specific generalization	Structure-aware molecular generation and scaffold modification
Transformer and molecular language models	Model chemical strings as sequences using autoregressive or conditional language modeling	SMILES, randomized SMILES, SELFIES, molecular tokens	Enabled scalable sequence-based generation and conditional property-guided design	Growing evidence for scalable generation and controllable molecule design	Output quality depends on training data, tokenization, conditioning relevance, and benchmark realism	Conditional molecule generation, target-informed design, and large-scale chemical language modeling
Diffusion models	Learn molecular distributions through iterative denoising or structure-conditioned generation	Molecular graphs, 3D molecular coordinates, geometric representations	Emerging approach for geometry-aware and structure-conditioned molecular design	Early-stage but rapidly expanding evidence base	Limited prospective validation and less mature benchmarking in drug discovery compared with RL or transformers	3D-aware molecular generation, structure-based design, and future geometry-constrained lead optimization
Hybrid synthesis-aware generators	Combine molecular generation with synthetic accessibility, retrosynthesis, or feasibility scoring	Graphs, SMILES, retrosynthetic route features, reaction-informed embeddings	Directly addresses the gap between valid molecules and makeable molecules	Promising but less consistently represented across the evidence base	Feasibility often remains post-hoc rather than embedded during generation	Translationally realistic candidate prioritization and design–make–test planning

Molecular Representations and Chemical Validity

Molecular representation was a major determinant of chemical validity, model behavior, and evaluation interpretation across the included studies. SMILES remained widely used because of its compatibility with sequence models, and randomized SMILES was reported to improve molecular generative modeling by augmenting string representations [27]. SELFIES was introduced as a robust molecular string representation designed to reduce invalid molecular outputs [26], while graph-based models represented atoms and bonds more directly [24, 25]. Several studies suggested that representation choice influenced

not only validity but also diversity, novelty, optimization behavior, and the ease of incorporating three-dimensional or synthetic constraints.

Synthetic Feasibility Assessment

Synthetic feasibility was assessed inconsistently across the reviewed studies, despite repeated recognition that generated molecules must be practical to make. SCScore provided a learned synthetic complexity metric based on reaction data [17], and RAscore introduced a rapid retrosynthesis-informed accessibility classifier for AI-driven retrosynthetic planning. Some generative pipelines incorporated synthetic accessibility into optimization or filtering, but several studies applied feasibility measures only after generation, limiting their influence on the learned design process. This pattern indicated that synthetic realism remains a major bottleneck between molecular generation and experimentally actionable drug discovery.

Multi-Objective Optimization and Property Control

Multi-objective optimization was common in studies that attempted to balance potency-related proxies with physicochemical, ADMET-like, and synthetic criteria. Reinforcement learning frameworks were frequently used for this purpose because reward functions could combine multiple molecular properties [7, 8, 21, 22]. Transformer and conditional generation studies also explored property-guided molecular design, although conditioning variables varied widely across studies [12, 28]. The evidence suggested that multi-objective optimization was methodologically mature in silico, but target engagement, assay relevance, and feasibility constraints were not consistently validated experimentally.

Pharmacological and Experimental Validation

Experimental pharmacological validation was rare relative to the volume of computational work. The DDR1 inhibitor study showed that deep learning could support rapid identification of potent kinase inhibitors with experimental follow-up, but comparable prospective validation was uncommon across the broader literature. Several studies reported computational target scores, docking, or predicted ADMET properties, yet these endpoints were usually not equivalent to biochemical, cellular, or in vivo validation. The evidence therefore supported a cautious interpretation: generative models can prioritize plausible candidates, but only a small subset of reported molecules has been tested in disease-relevant experimental systems.

Evaluation Metrics and Benchmarks Used

Evaluation metrics were dominated by validity, uniqueness, novelty, diversity, property distribution, and goal-directed optimization scores. MoleculeNet provided a broader benchmark foundation for molecular machine learning [13], while GuacaMol and MOSES specifically addressed benchmarking of de novo molecular generation models [14, 15]. The Fréchet ChemNet Distance added a distributional metric for comparing generated molecules with reference chemical sets [16]. Although these benchmarks improved comparability, they did not fully capture synthesis feasibility, biological novelty, assay translation, or developability.

Barriers to Translation

Several recurring barriers limited translation from in silico molecular generation to lead-like compounds suitable for development. Synthetic feasibility remained a central barrier because models often generated molecules that were chemically valid but not necessarily synthetically practical [17]. Generalization to novel targets was also uncertain, particularly when models were trained or optimized on narrow datasets or benchmark-specific objectives [13-15]. Finally, the scarcity of prospective pharmacological validation meant that translational readiness was usually inferred rather than demonstrated.

The Proliferation of Generative Architectures

The reviewed literature showed rapid proliferation of generative architectures, from early VAEs and reinforcement learning systems to transformers, graph models, adversarial models, and diffusion approaches. This diversity has expanded the technical repertoire of de novo molecular design [1, 3], but it has also produced fragmented reporting practices and limited comparability across studies. Several reviews have emphasized that generative drug design is now technically sophisticated [2, 4, 5, 18, 19], yet the practical value of increasingly complex models remains difficult to judge without consistent validation. The evidence therefore suggests that architectural novelty alone is insufficient as a marker of progress.

Figure 2 synthesizes the review findings into an evidence-to-translation map showing how rapid model innovation narrows at the stages of synthesis feasibility, pharmacological validation, and closed-loop drug discovery.

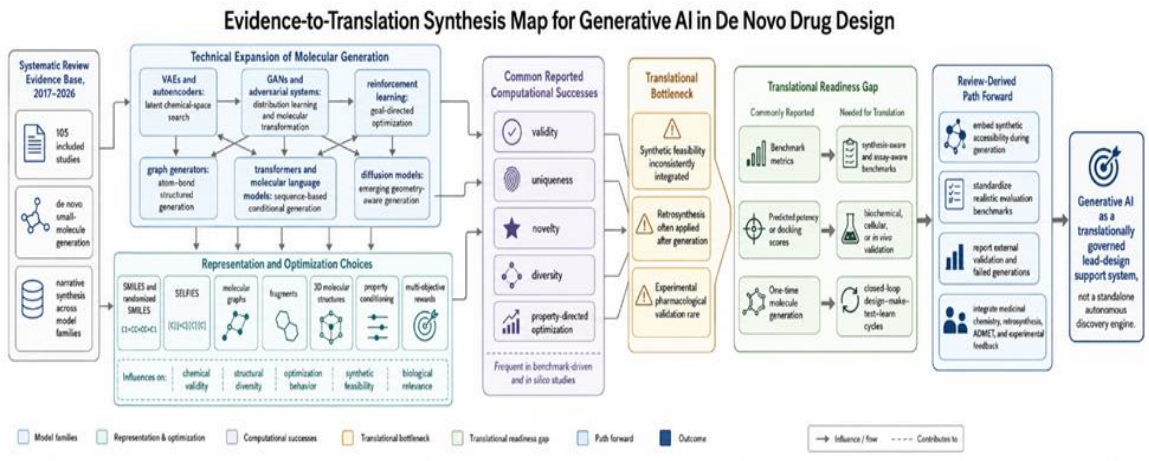

Figure 2. Evidence-to-Translation Synthesis Map for Generative AI in De Novo Drug Design

Validity and Novelty Are High, but Synthetic Feasibility Is Ignored

Many studies reported generation of valid and novel molecules, but these outputs often remained disconnected from practical synthesis. Benchmarks such as GuacaMol and MOSES encouraged standardized evaluation of chemical validity and distributional quality [14, 15], while the Fréchet ChemNet Distance provided an additional distributional similarity measure [16]. However, synthetic complexity and retrosynthesis-based feasibility tools such as SCScore and RAscore were not uniformly integrated into generation objectives [17]. As a result, several models appeared successful by benchmark metrics while still leaving unresolved whether generated compounds could be made efficiently.

The Missing Pharmacological Validation

The most important translational weakness in the evidence base was the limited frequency of pharmacological validation. The DDR1 inhibitor study demonstrated that AI-generated molecules can enter experimental testing pipelines, but this type of prospective validation remained exceptional rather than routine. Many studies relied on predicted properties, docking scores, or benchmark objectives, which can be useful for triage but do not establish target engagement or biological activity. This gap suggests that generative AI has advanced faster as a computational methodology than as an experimentally validated drug discovery workflow.

Evaluation Benchmarks Are Not Yet Fit-for-Purpose

Current evaluation benchmarks have improved reproducibility, but they do not yet fully represent the requirements of drug discovery. MoleculeNet, GuacaMol, and MOSES provided important shared tasks and datasets [13-15], yet they primarily evaluated computational behavior rather than end-to-end development potential. Metrics such as validity, uniqueness, novelty, and diversity are necessary but insufficient because they do not measure synthesis route availability, assay feasibility, safety liabilities, or target-specific pharmacology [16, 17, 29]. The reviewed evidence therefore indicates a need for benchmarks that reward realistic, synthesis-aware, and experimentally grounded molecular design.

Table 2 provides an evaluation-to-translation framework showing why common molecular-generation metrics are necessary for technical assessment but insufficient for judging experimental and drug discovery readiness.

Table 2. Evaluation-to-Translation Framework for Generative AI Drug Design Studies

Evaluation domain	Commonly reported indicator	What the indicator demonstrates	What the indicator does not demonstrate	Translationally stronger evidence standard	Interpretation for this review
Chemical validity	Percentage of generated molecules with valid chemical syntax or graph structure	The model can produce chemically interpretable outputs	Synthetic feasibility, biological activity, novelty beyond training data, or developability	Validity combined with novelty, diversity, synthesizability, and external validation	Necessary but insufficient; validity is a minimum technical requirement, not evidence of drug discovery impact
Uniqueness and novelty	Fraction of non-duplicate molecules and molecules absent from the training set	The model can generate non-redundant and previously unseen structures	Whether molecules are useful, makeable, potent, safe, or patent-relevant	Novel structures filtered through medicinal chemistry, retrosynthesis, and biological relevance criteria	Frequently reported but often overinterpreted as innovation without downstream validation
Diversity and distributional similarity	Internal diversity, scaffold diversity, Fréchet ChemNet Distance, or	The model samples broadly or resembles reference chemical distributions	Whether generated molecules solve a target-specific drug discovery problem	Diversity evaluated alongside target activity, ADMET constraints, and	Useful for comparing generators but weak as a standalone translational endpoint

	benchmark distribution scores			synthesis route availability	
Property optimization	Predicted logP, QED, docking score, target score, or multi-objective reward	The model can steer generation toward predefined computational objectives	Whether optimized molecules bind experimentally, enter cells, avoid toxicity, or remain synthesizable	Multi-parameter optimization with experimentally relevant assays and feasibility constraints	Strong in silico performance may reflect reward design rather than real pharmacological value
Synthetic accessibility	Synthetic accessibility score, SCScore, RAscore, retrosynthetic route prediction	The molecule may be easier or harder to synthesize computationally	Actual route success, yield, cost, scalability, purification burden, or chemical stability	Retrosynthesis-informed prioritization plus expert medicinal chemistry review and attempted synthesis	Under-integrated across the literature; often applied after generation rather than during model optimization
Benchmark performance	GuacaMol, MOSES, MoleculeNet-related tasks, or related standardized comparisons	The model performs comparably on shared tasks	End-to-end drug discovery readiness, assay relevance, synthesis feasibility, or generalization to new targets	Benchmarks that incorporate retrosynthesis, purchasable building blocks, assay-ready selection, and external validation	Current benchmarks improve reproducibility but remain incomplete proxies for translational success
Pharmacological validation	Docking, predicted bioactivity, biochemical assay, cellular assay, in vivo testing	The model's outputs may have target relevance or biological activity	Docking and predicted scores alone do not establish activity or developability	Prospective biochemical or cellular testing, followed by iterative design refinement	Rare in the reviewed evidence; the strongest translational gap identified by the review
Closed-loop learning	Iterative design–make–test–learn cycles with feedback into the model	The system can improve through experimental evidence	One-time computational ranking does not prove learning from real outcomes	Repeated experimental cycles with synthesis, assay feedback, model updating, and transparent reporting	Largely absent; represents a major future research priority

Cross-Cutting Challenges: Data Quality, Generalization, and Reproducibility

Data quality, generalization, and reproducibility issues were shared across model families. Sequence-based models depended on representation quality and training set coverage [12, 26, 27], while graph and adversarial models were sensitive to molecular graph construction, optimization settings, and distributional assumptions [9, 24, 25]. Reinforcement learning systems were particularly vulnerable to reward hacking when objectives were poorly aligned with realistic drug discovery priorities [7, 8, 21, 22]. Across these categories, incomplete reporting of negative results, failed generations, and external validation reduced confidence in generalizability.

Comparison with Traditional De Novo Design

Generative AI differs from traditional de novo design by learning molecular distributions and proposing novel structures through data-driven optimization rather than relying mainly on rule-based enumeration or hand-designed transformations. Several studies suggested that deep generative systems can explore chemical space flexibly [6-8], and transformer or graph models further expanded this capacity [12, 24]. However, traditional medicinal chemistry workflows often incorporate synthesis planning and experimental feedback more explicitly than many generative AI studies. The evidence therefore suggests that generative models may outperform older approaches in novelty generation but have not consistently surpassed them in practical lead delivery.

Toward Realistic De Novo Design

A realistic path forward requires closer integration of molecular generation with synthesis planning, pharmacological testing, and iterative feedback. Synthetic accessibility methods such as SCScore and RAscore provide components for feasibility-aware design [17, 29], while integrated approaches have begun to connect AI-based generation with synthetic accessibility constraints. Reinforcement learning and conditional generation frameworks may support closed-loop optimization if reward functions include experimental and retrosynthetic evidence rather than only predicted properties [21, 22, 28]. The reviewed literature therefore points toward hybrid pipelines in which generative models propose candidates, retrosynthesis filters prioritize feasible structures, and experimental assays refine subsequent design cycles.

Limitations

Review Limitations

This review was limited by English-language inclusion, possible publication bias toward positive generative-model results, and substantial heterogeneity across datasets, architectures, and validation endpoints. Because included studies used different representations, benchmarks, targets, and feasibility criteria, quantitative meta-analysis was not appropriate [13-16]. Some relevant conference papers or preprints may have contributed to the field but were not prioritized unless peer-reviewed and

sufficiently detailed. The synthesis therefore emphasizes systematic patterns rather than pooled comparative performance estimates, consistent with prior narrative reviews of generative molecular design [1-5, 18-20].

Evidence Base Limitations

The evidence base itself was dominated by methodological studies reporting in silico generation, benchmark performance, and predicted molecular properties. Studies integrating synthetic feasibility during generation were less common, despite the availability of tools such as SCScore, RAscore, and synthesis-aware evaluation approaches [17, 29]. Experimental validation was especially sparse, with the DDR1 inhibitor study standing out as a prominent example rather than a typical design standard. These limitations mean that conclusions about translational readiness should be interpreted cautiously, because most reported molecules have not been synthesized, assayed, or optimized in realistic drug discovery programs.

Comparison With Prior Reviews

Prior reviews have mapped important parts of the generative AI drug design landscape, but many focused on selected model classes, specific representation strategies, or broad conceptual overviews. Reviews of deep generative molecular design described the growth of VAEs, GANs, and reinforcement learning [3, 4], while other surveys emphasized the broader role of generative AI in molecular and protein generation [1]. Additional reviews highlighted de novo molecular design as a maturing computational field but did not uniformly assess whether generated compounds were synthetically feasible or experimentally validated [2, 18]. As a result, the earlier review literature established the technical breadth of the field but left unresolved questions about translational maturity.

This review differs from prior reviews by systematically spanning VAEs, GANs, reinforcement learning, diffusion models, transformers, molecular representations, synthesis-aware scoring, benchmarking, and pharmacological validation. Earlier discussions of SELFIES, randomized SMILES, and graph representations clarified how molecular encoding affects generation [24-27], while benchmark-focused papers provided a shared language for evaluating generated molecules [13-16]. This review integrates those technical strands with synthesis-focused methods such as SCScore and RAscore [17, 29]. It therefore evaluates not only whether models generate plausible molecules but also whether those molecules are likely to become experimentally actionable candidates.

The review also emphasizes that the translational gap is not a secondary concern but a central limitation of the field. Studies such as REINVENT and MolGPT showed how optimization and language modeling can generate property-guided molecules [12, 21, 22], yet these advances do not automatically imply biological efficacy or developability. The DDR1 inhibitor study remains an important example of experimental follow-up, but such prospective validation was not widely reproduced across model families. Compared with prior reviews, this synthesis places greater weight on the pathway from molecular generation to synthesis, testing, and lead-like progression [5].

Recommendations

For Researchers

Researchers should report synthetic accessibility, external validation, and prospective testing more consistently in de novo molecular design studies. Standard molecular-generation metrics such as validity, uniqueness, novelty, and diversity should be retained [14, 15], but they should be supplemented with feasibility-aware measures such as SCScore, retrosynthesis-based accessibility, or related synthesis planning outputs [17, 29]. Reward functions and conditioning objectives should avoid over-optimizing narrow proxies without biological relevance, a concern especially relevant to reinforcement learning systems [7, 8, 26, 27]. Where possible, generated molecules should be prioritized through closed-loop experimental feedback rather than only through predicted potency or docking-like scores.

For Journal Editors

Journal editors should require de novo design manuscripts to report whether generated molecules were assessed for synthetic feasibility, novelty relative to training data, and biological relevance. Studies using standard benchmarks such as MoleculeNet, GuacaMol, or MOSES should clearly describe dataset splits, benchmark configuration, and whether optimization objectives could create data leakage or unrealistic enrichment [13-15]. Manuscripts claiming drug discovery relevance should distinguish computational molecular generation from validated pharmacological discovery, especially when no synthesis or assay data are provided [5]. Editorial standards could also encourage deposition of generated molecules, code, training data descriptions, and failed-generation analyses to strengthen reproducibility.

For the Community

The community should develop standardized benchmarks that include synthesis feasibility, multi-parameter optimization, and target-specific validation rather than relying only on chemical plausibility metrics. Existing benchmarks have been valuable for comparability [14, 15], and distributional metrics such as the Fréchet ChemNet Distance have added useful evaluation dimensions [16]. However, future benchmarks should incorporate practical constraints such as retrosynthetic accessibility, purchasable building blocks, assay-ready compound selection, and ADMET-informed prioritization [17]. Such benchmarks would better reflect how generative models are expected to function in real drug discovery environments.

Research Gaps

Prospective, Closed-Loop Validation

A major gap is the absence of broadly demonstrated fully closed-loop de novo design-make-test cycles across diverse targets. Although reinforcement learning and conditional generation models can iteratively optimize molecules in silico [7, 8, 21, 22, 28], most studies do not close the loop with synthesis, assay feedback, and model retraining. The DDR1 inhibitor report provides evidence that rapid AI-supported hit identification can be experimentally pursued, but it remains an uncommon example rather than a routine workflow. Future research should test whether generative systems can improve after each experimental cycle rather than merely generate candidates for one-time computational ranking.

Robust Synthetic Feasibility Scoring in Generative Loops

Synthetic feasibility scoring remains under-integrated into generative pipelines. SCScore and RAscore provide practical methods for estimating synthetic complexity and retrosynthetic accessibility [17, 29], and synthesis-aware generative design has begun to address this limitation. However, many studies still treat feasibility as a post-generation screen rather than as a constraint that shapes molecular generation from the beginning. Research is needed to determine whether incorporating synthesis feasibility during training, reinforcement learning, or conditional sampling improves the fraction of generated compounds that can be synthesized without sacrificing novelty or potency-related objectives.

Implications

For Research Practice

The field should move from demonstrating algorithmic novelty toward demonstrating translational utility. Early latent-space, adversarial, reinforcement learning, transformer, and graph-based approaches established that generative systems can produce novel chemical structures [6-10, 12, 21-25]. The next stage should prioritize whether generated compounds are synthesizable, testable, and biologically meaningful under realistic constraints [17, 29]. This shift would align model development with drug discovery practice rather than with isolated benchmark optimization.

For Drug Discovery

Generative models are currently best positioned as tools for hit expansion, scaffold exploration, and lead optimization rather than fully autonomous ab initio drug discovery systems. Reinforcement learning and conditional transformer models can guide molecules toward desired property profiles [7, 8, 21, 22, 28], and representation advances such as SELFIES and graph generation can improve chemical robustness [24-26]. However, the limited integration of prospective pharmacology and synthesis planning means that human medicinal chemistry judgment remains central. In practical pipelines, generative AI may be most useful when embedded within multidisciplinary workflows that include retrosynthesis, ADMET assessment, and experimental testing [17, 29].

Conclusion

Generative AI for de novo drug design has diversified rapidly and can generate valid, novel, and property-directed molecules across multiple model families. The field now includes latent-variable models, adversarial systems, reinforcement learning, molecular language models, graph generators, and emerging diffusion approaches.

The literature remains dominated by in silico methodology studies, and only a small, fragmented subset reports synthesis or pharmacological testing. As a result, the current evidence supports technical promise more strongly than demonstrated translational impact.

Future progress depends on closing the loop between molecular generation, synthetic planning, compound synthesis, biological testing, and model refinement. Studies that integrate these stages will provide more meaningful evidence than studies that report only computational novelty or benchmark performance.

A new generation of generative drug design research should emphasize translational readiness alongside algorithmic innovation. Synthetic feasibility, biological validation, transparent reporting, and realistic benchmarks should become standard expectations rather than optional additions.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Tang X, Dai H, Knight E, Wu F, Li Y, Li T, et al. A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. Brief Bioinform. 2024;25(4):bbae338.
2. Meyers J, Fabian B, Brown N. De Novo Molecular Design and Generative Models. Drug Discov Today. 2021;26(11):2707-15.
3. Cheng Y, Gong Y, Liu Y, Song B, Zou Q. Molecular design in drug discovery: a comprehensive review of deep generative models. Brief Bioinform. 2021;22(6):bbab344.
4. Tong X, Liu X, Tan X, Li X, Jiang J, Xiong Z, et al. Generative models for de novo drug design. J Med Chem. 2021;64(19):14011-27.
5. Zeng X, Wang F, Luo Y, Kang SG, Tang J, Lightstone FC, et al. Deep generative molecular design reshapes drug discovery. Cell Rep Med. 2022;3(12).
6. 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.
7. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminform. 2017;9(1):48.
8. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Sci Adv. 2018;4(7):eaap7885.
9. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Mol Pharm. 2017;14(9):3098-104.
10. Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, et al. Reinforced adversarial neural computer for de novo molecular design. J Chem Inf Model. 2018;58(6):1194-204.
11. Alakhdar A, Poczos B, Washburn N. Diffusion models in de novo drug design. J Chem Inf Model. 2024;64(19):7238-56.
12. Bagal V, Aggarwal R, Vinod PK, Priyakumar UD. MolGPT: molecular generation using a transformer-decoder model. J Chem Inf Model. 2021;62(9):2064-76.
13. 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.
14. Brown N, Fiscato M, Segler MH, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096-108.
15. 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.
16. Preuer K, Renz P, Unterthiner T, Hochreiter S, Klambauer G. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. J Chem Inf Model. 2018;58(9):1736-41.
17. Coley CW, Rogers L, Green WH, Jensen KF. SCScore: synthetic complexity learned from a reaction corpus. J Chem Inf Model. 2018;58(2):252-61.
18. Bilodeau C, Jin W, Jaakkola T, Barzilay R, Jensen KF. Generative models for molecular discovery: Recent advances and challenges. WIREs Comput Mol Sci. 2022;12(5):e1608.
19. Xu Y, Lin K, Wang S, Wang L, Cai C, Song C, et al. Deep learning for molecular generation. Future Med Chem. 2019;11(6):567-97.
20. Anstine DM, Isayev O. Generative models as an emerging paradigm in the chemical sciences. J Am Chem Soc. 2023;145(16):8736-50.
21. Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, et al. REINVENT 2.0: an AI tool for de novo drug design. J Chem Inf Model. 2020;60(12):5918-22.
22. 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.
23. Maziarka Ł, Pocha A, Kaczmarczyk J, Rataj K, Danel T, Warchoł M. Mol-CycleGAN: a generative model for molecular optimization. J Cheminform. 2020;12(1):2.
24. Samanta B, De A, Jana G, Gómez V, Chattaraj P, Ganguly N, et al. Nevae: A deep generative model for molecular graphs. J Mach Learn Res. 2020;21(114):1-33.
25. Kwon Y, Yoo J, Choi YS, Kang S. Efficient learning of non-autoregressive graph variational autoencoders for molecular graph generation. J Cheminform. 2019;11:21.
26. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach Learn: Sci Technol. 2020;1(4):045024.
27. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond JL, et al. Randomized SMILES strings improve the quality of molecular generative models. J Cheminform. 2019;11(1):71.
28. Wang Y, Zhao H, Sciabola S, Wang W. cMolGPT: a conditional generative pre-trained transformer for target-specific de novo molecular generation. Molecules. 2023;28(11):4430.
29. 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.