Volume 1, Issue 2, July-September 2026

e-ISSN: AWAITING

INTERNATIONAL JOURNAL OF
Medical and Healthcare Sciences (IJMedHS)

OPEN ACCESS

Journal homepage: https://jmedhsscience.com/

Generative Artificial Intelligence for De Novo Small-Molecule Drug Discovery: an Innovative Strategy in Modern Drug Development

Sangram Jayvant Mali[1*], Anuj Narendra Umare[2]

[1,2]Bharati Vidyapeeth College of Pharmacy, Kolhapur
[1]sangrammali2448@gmail.com, [2]umareanuj89@gamil.com

Corresponding Auther:- sangrammali2448@gmail.com

Article Info

Received: 28 July 2026
Accepted: 30 August 2026
Published: 02 September 2026

*Mali, S. J.,
Generative Artificial Intelligence for De Novo Small-Molecule Drug Discovery: an Innovative Strategy in Modern Drug Development, *International Journal of Medical & Healthcare Sciences [IJMedHS], 1(2), 09-17*
DOI:

ABSTRACT

Modern drug discovery is limited by the enormous size of chemical space, the high cost of experimental screening, and the frequent failure of promising candidates during development. Generative artificial intelligence has emerged as an innovative strategy for de novo small-molecule drug discovery because it can learn relationships between molecular structures and desired biological or physicochemical properties and then propose previously unreported molecular structures. Unlike conventional screening, which primarily searches existing chemical libraries, generative approaches can explore new regions of chemical space while applying constraints related to potency, selectivity, physicochemical behavior and synthetic feasibility. The workflow begins with high-quality chemical and biological data, followed by molecular representation, model training, generation of candidate structures, multi-property filtering, synthetic assessment and experimental validation. Recurrent neural networks, variational approaches, generative adversarial approaches, graph-based models and transformer-based architectures have been investigated for molecular generation. Early studies demonstrated the ability of generative systems to produce focused libraries and biologically active compounds, while more recent work has progressed to candidates entering human clinical trials. The discovery and development of rentosertib, a small-molecule inhibitor of TRAF2- and NCK-interacting kinase for idiopathic pulmonary fibrosis, provides an important translational example of an artificial-intelligence-driven program progressing from target discovery and molecule generation through preclinical evaluation and phase 2a clinical testing. Nevertheless, generative artificial intelligence does not eliminate the need for medicinal chemistry, pharmacology, toxicology and clinical evidence. Data quality, model bias, chemical validity, synthetic accessibility, uncertainty, interpretability, reproducibility and regulatory acceptance remain important challenges. Properly integrated with experimental science and human expertise, generative artificial intelligence can function as a powerful hypothesis-generation and candidate-prioritization strategy for modern drug development.

Keywords: Generative artificial intelligence; De novo drug discovery; Molecular generation; Small-molecule design; Machine learning; Chemical space; Molecular optimization; Drug development; Computational drug discovery; Artificial intelligence.

1. Introduction

Drug discovery is a complex and high-risk process in which biological understanding, medicinal chemistry, pharmacology, toxicology and clinical research must converge to produce a safe and effective medicine. Traditional discovery relies heavily on target selection, screening of available chemical collections, iterative medicinal chemistry and repeated biological testing.

These activities can require substantial time and resources because most screened compounds do not possess the required combination of potency, selectivity, pharmacokinetic behavior, safety and developability. Machine learning has therefore attracted attention as a means of extracting useful patterns from large biological and chemical datasets and supporting decision making across the discovery pipeline [1,2].

Generative artificial intelligence represents a more specific development within this broader computational field. Instead of only predicting a property of a molecule that already exists, a generative model can propose new molecular structures according to learned patterns and user-defined objectives. The approach treats molecular design as a constrained generation problem. A model is trained on known chemical structures and associated information and subsequently samples or constructs candidate molecules that may possess desired characteristics. Generative models can therefore support exploration of chemical space beyond the compounds directly represented in the training library [3-5].

The importance of this approach arises from the enormous theoretical chemical space available for drug-like molecules. Experimental screening can examine only a small fraction of this space. Computational generation can reduce the search space by proposing candidates that satisfy multiple criteria before laboratory synthesis. Early studies using recurrent neural networks demonstrated that molecular generators could produce focused libraries enriched for structures associated with desired biological activity [6]. Deep generative approaches subsequently demonstrated the ability to optimize multiple molecular objectives, including novelty, biological activity and synthetic feasibility [7].

Generative artificial intelligence should not be regarded as an autonomous replacement for medicinal chemistry. Its practical value lies in creating hypotheses and prioritizing compounds for experimental testing. The model output must be evaluated for chemical validity, stability, synthetic accessibility, activity, selectivity, absorption, distribution, metabolism, excretion and toxicity. The importance of this human and experimental layer has become increasingly clear as the field has matured. Recent assessments indicate that the number of impressive computational demonstrations has grown faster than evidence of clinically meaningful impact, making prospective validation and translational relevance essential [8].

Advantages

Generative artificial intelligence can explore chemical space rapidly, produce novel molecular scaffolds, optimize several properties simultaneously and reduce the number of compounds requiring synthesis. It can also incorporate structural information, biological activity and physicochemical constraints during candidate generation. In appropriate settings, this can shorten the design-make-test cycle and help medicinal chemists focus on a smaller set of experimentally useful candidates [1,3].

Challenges

The principal limitations are data quality, limited and biased training datasets, poor representation of chemical context, invalid or unstable generated structures, difficulty predicting biological activity outside the training distribution, limited interpretability and uncertainty in experimental translation. A model can generate a molecule that is mathematically attractive but chemically impractical or biologically inactive. Benchmarking has also shown that data scarcity and highly imbalanced datasets remain difficult problems for molecular machine learning [5].

Factors affecting performance

The success of a generative system depends on the quality and relevance of training data, molecular representation, model architecture, objective functions, diversity constraints, synthetic accessibility, validation strategy and the quality of experimental feedback. Data provenance is particularly important because duplicated measurements, inconsistent assay conditions and publication bias can produce apparently strong models that fail prospectively. Model evaluation should therefore include external validation, scaffold-based splitting where appropriate, uncertainty assessment and prospective experimental testing [1,5,8].

Introduction of The Generative Artificial Intelligence Technique

Generative artificial intelligence for drug discovery is a computational strategy in which a trained model learns the statistical or structural relationships present in chemical and biological data and then generates new molecular candidates. The technique differs from simple molecular property prediction because its principal output is a new chemical structure rather than only a predicted numerical value.

Molecular representation

A molecular generator requires a representation that can be processed computationally. Common representations include molecular strings, molecular graphs, fingerprints and three-dimensional structural descriptions. String-based approaches represent a molecule as a sequence and can use recurrent or transformer architectures. Graph-based approaches represent atoms as nodes and chemical bonds as edges, allowing the model to learn structural relationships directly. The choice of representation affects validity, diversity, computational efficiency and the ability of the model to capture chemically meaningful relationships [5,6].

Training data

Training data may contain chemical structures linked with measured biological activity, physicochemical properties, toxicity information, synthetic reactions or structural information. Public resources provide large collections of curated bioactivity and chemical data, but the usefulness of these resources depends on annotation quality and

experimental consistency. The ChEMBL database, for example, contains millions of compounds and bioactivity measurements collected from publications, patents and deposited datasets, making it an important source for data-driven drug discovery [17]. DrugBank provides complementary information on drugs, targets and pharmacological relationships [18].

Model training

During training, the generative model learns a probability distribution or latent representation of the molecular data. The objective may be to reproduce chemically valid structures, generate molecules with selected properties, or optimize several objectives simultaneously. The model is then used to generate candidate molecules, which are filtered according to predefined criteria.

Generation and optimization

A practical workflow usually combines generation with property optimization. A candidate can be evaluated for predicted biological activity, physicochemical characteristics, novelty, structural alerts and synthetic feasibility. Molecules that fail predefined requirements are discarded, while promising structures are ranked for synthesis. This iterative process can be repeated after experimental results are obtained.

Table 1. General workflow of generative artificial intelligence-based de novo drug discovery

Stage	Major activity	Expected output
Data preparation	Collection, cleaning and standardization of chemical and biological data	Curated training dataset
Molecular representation	Encoding molecules as strings, graphs or other machine-readable structures	Computational molecular representation
Model development	Training and validation of the generative model	Validated molecular generator
Molecular generation	Creation of new candidate structures	Novel molecular library
Multi-property filtering	Assessment of activity, selectivity, physicochemical properties and developability	Prioritized candidates
Synthetic assessment	Evaluation of synthetic accessibility and practical chemistry	Synthesizable candidates
Experimental validation	Chemical synthesis and biological testing	Confirmed active compounds
Iterative optimization	Feeding experimental observations back into the design process	Improved candidates

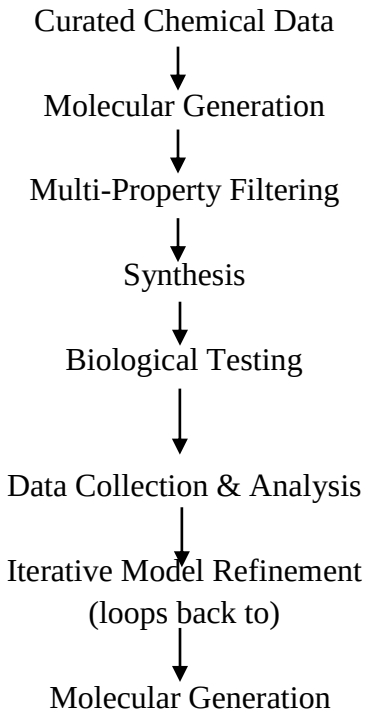

Figure 1. Suggested original schematic showing the workflow from curated chemical data to molecular generation, multi-property filtering, synthesis, biological testing and iterative model refinement.

Generation of Candidate Molecules

Generative models can be designed to produce molecules sequentially, construct molecular graphs, or operate through latent representations. Recurrent neural networks demonstrated an early ability to generate focused molecular libraries by learning chemical sequences and fine-tuning generation toward molecules associated with selected biological targets [6]. Later approaches used deep generative models to balance novelty, biological activity and synthetic feasibility. A notable example is the generative tensorial reinforcement learning system used to identify inhibitors of discoidin domain receptor 1. The study reported active compounds within a short computational and experimental cycle, demonstrating how generative design can narrow the search for new chemical matter [7].

A major advantage of generative design is multi-objective optimization. Drug candidates rarely need only high potency. They may also require acceptable solubility, lipophilicity, molecular size, metabolic stability, selectivity and synthetic accessibility. A generation strategy can therefore be configured to reward several desirable properties and penalize undesirable characteristics. This concept moves computational design closer to the real decision-making problem faced by medicinal chemists [3,4].

Synthetic feasibility is especially important. A generated molecule has little practical value if it cannot be synthesized efficiently. Data-driven reaction prediction has consequently become an important supporting capability. Transformer-based reaction prediction has demonstrated high accuracy in predicting chemical transformations and can provide uncertainty estimates for predicted reactions [9]. In a complete workflow, generated structures can therefore be subjected to synthetic route analysis before being selected for laboratory preparation.

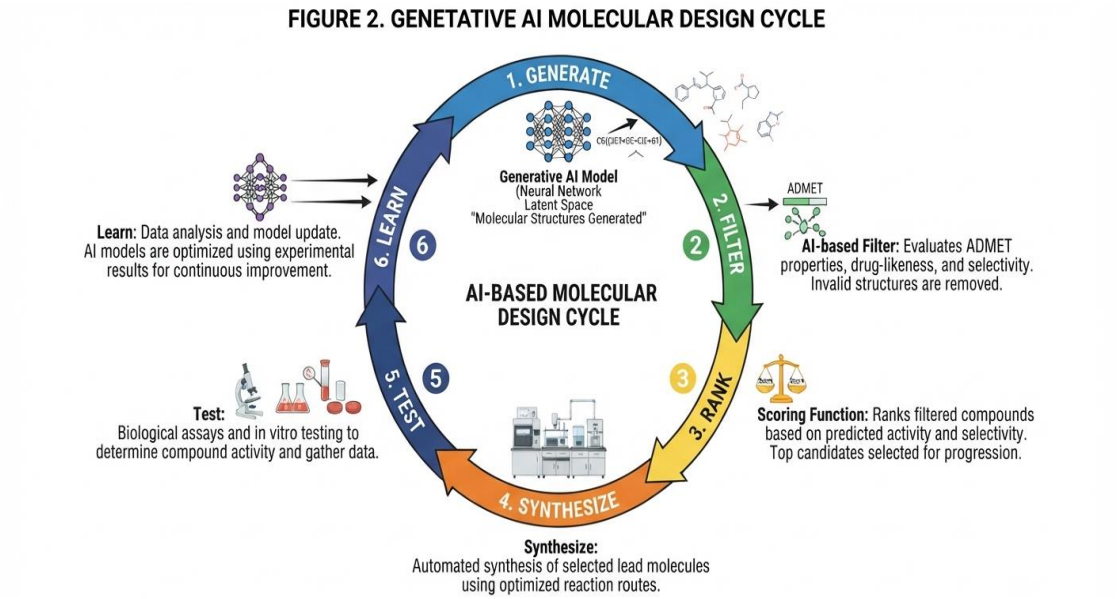

Figure 2. Suggested original schematic showing the generate-filter-rank-synthesize-test-learn cycle of generative artificial intelligence-based molecular design.

In Vitro Evaluation

In vitro evaluation is essential because computational predictions are hypotheses rather than proof of biological activity. The first experimental stage is normally confirmation that the synthesized compound is chemically correct and sufficiently pure. The biological evaluation can then progress from target-level testing to cellular systems.

Target-level evaluation

For a target-directed program, biochemical assays can determine whether the generated compound interacts with the intended protein or biological target. Depending on the mechanism, suitable assays may measure enzyme inhibition, receptor activation or inhibition, ligand displacement, binding affinity or another direct target response. Concentration-response experiments are used to determine potency and establish whether the predicted activity is experimentally reproducible.

Cell-based evaluation

A compound that shows target-level activity should be tested in relevant cells. Cell-based assays can determine whether the compound crosses cellular barriers, reaches

the intracellular target and produces the expected pharmacological response. Cytotoxicity toward relevant non-target cells should also be evaluated. Where possible, biomarker analysis can connect the molecular target to the downstream biological response.

Selectivity and mechanism

Selectivity testing is important because strong activity against an intended target does not automatically imply therapeutic usefulness. A candidate may interact with related proteins or unrelated cellular targets. Orthogonal assays, target-engagement measurements and mechanistic biomarkers can strengthen evidence for the proposed mode of action.

Absorption, distribution, metabolism and excretion-related in vitro evaluation

Early assessment of absorption, distribution, metabolism and excretion-related characteristics can identify candidates with poor developability before expensive animal studies. Common studies include aqueous solubility, membrane permeability, microsomal stability, plasma stability, protein binding and metabolic interaction assessment. Early toxicity screening can identify cytotoxicity, genotoxicity or other liabilities.

Characterization

Characterization confirms the identity, purity and key physicochemical properties of synthesized candidates. Although generative artificial intelligence may propose the molecular structure, the experimentally prepared compound must be independently characterized.

Solubility and physicochemical properties

Aqueous solubility is important because insufficient solubility can limit exposure and formulation options. Partition behavior, ionization characteristics and other physicochemical properties influence permeability, distribution and formulation requirements. These properties should be measured or experimentally confirmed rather than accepted solely from computational predictions.

Spectroscopic characterization

Infrared spectroscopy can support identification of characteristic functional groups. Nuclear magnetic resonance spectroscopy provides structural information about the molecular environment of atoms and is particularly useful for confirming connectivity and stereochemical consistency when applicable. Mass spectrometry can verify molecular mass and support impurity assessment. Chromatographic methods can determine chemical purity and detect degradation products.

Thermal and solid-state characterization

Melting behavior, crystallinity and solid-state form may influence stability, solubility and formulation performance. When relevant, differential scanning calorimetry and powder X-ray diffraction can be used to characterize thermal and crystalline behavior.

Table 2. Experimental characterization of an artificial-intelligence-generated drug candidate

Parameter	Typical purpose
Appearance	Preliminary physical assessment
Solubility	Assessment of aqueous and formulation-relevant behavior
Melting behavior	Identity and solid-state assessment
Infrared spectroscopy	Functional-group confirmation
Nuclear magnetic resonance spectroscopy	Structural confirmation
Mass spectrometry	Molecular-mass confirmation
Chromatographic purity	Quantification of chemical purity and related substances
Solid-state analysis	Assessment of crystallinity and polymorphic behavior

Preclinical Trials

After successful in vitro validation, promising candidates undergo preclinical development. The objective is to establish pharmacological activity, exposure, safety and an initial therapeutic window before human testing. Preclinical studies normally integrate pharmacokinetics, pharmacodynamics, efficacy and toxicology.

Pharmacokinetic evaluation

Pharmacokinetic studies determine how the candidate behaves after administration. Important parameters include systemic exposure, maximum concentration, time to maximum concentration, elimination behavior and bioavailability. The route of administration is selected according to the intended clinical use.

Pharmacodynamic evaluation

Pharmacodynamic studies establish the relationship between drug exposure and biological response. Biomarkers can provide evidence that the intended target is modulated in vivo. Linking exposure, target engagement and therapeutic response is particularly important for computationally designed molecules because it tests the complete mechanistic hypothesis.

Efficacy studies
Disease-relevant animal models can be used to determine whether the candidate produces a meaningful therapeutic effect. Model selection should be justified by disease biology and the intended mechanism. A positive result should be interpreted together with exposure and pharmacodynamic evidence rather than as an isolated efficacy signal.

Safety and toxicology
Safety assessment includes acute and repeated-dose toxicity, clinical observations, clinical pathology, organ-specific assessment and other studies appropriate to the candidate and intended clinical development plan. Genotoxicity, safety pharmacology, reproductive toxicity and other investigations are performed according to development stage and regulatory requirements.
A notable translational example is INS018_055, later known as rentosertib. An artificial-intelligence-driven program identified TRAF2- and NCK-interacting kinase as an anti-fibrotic target and generated a small-molecule inhibitor. The candidate demonstrated anti-fibrotic activity in multiple experimental models, and its safety, tolerability and pharmacokinetic behavior were subsequently evaluated in phase 0 and phase I studies. The reported period from target discovery to preclinical candidate nomination was approximately 18 months [19].

Clinical Trials
The purpose of clinical development is to determine whether the candidate is safe, pharmacologically active and clinically beneficial in humans. Artificial intelligence can support candidate selection and trial planning, but it cannot replace clinical evidence.

Phase 0 and phase I
Early clinical studies evaluate pharmacokinetics, tolerability, dose escalation and initial safety. For computationally generated candidates, these studies provide the first opportunity to test whether model-derived expectations regarding exposure and biological behavior translate to humans.

Phase II
Phase II studies provide preliminary evidence of efficacy in patients while continuing safety assessment. Trial design should include clinically meaningful endpoints and appropriate controls. The example of rentosertib is particularly relevant because a randomized phase 2a study in patients with idiopathic pulmonary fibrosis reported safety outcomes and exploratory efficacy findings after 12 weeks of treatment [20].

Phase III
A promising phase II result must be confirmed in larger and appropriately powered studies. Phase III trials determine whether the benefit-risk profile is sufficiently robust for regulatory submission. Artificial intelligence does not reduce the evidentiary requirements of clinical development.

Regulatory considerations
Regulatory evaluation requires confidence in the quality, traceability and reliability of data used to support decisions. The United States Food and Drug Administration has highlighted issues including data quality, model credibility, data sharing and cybersecurity in discussions of artificial intelligence and machine learning applications in drug development [16]. Consequently, model development should be documented, datasets should be traceable, changes to models should be controlled, and computational predictions should be linked clearly to experimental evidence.

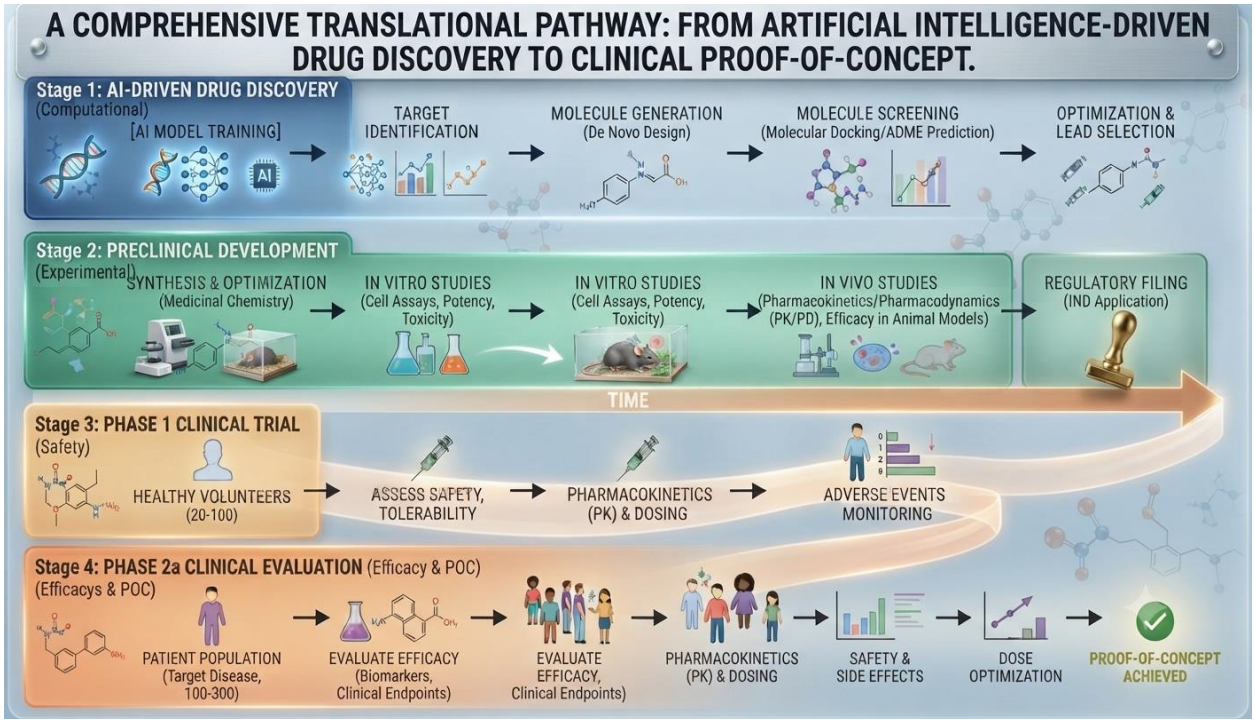

Figure 3. Suggested original schematic showing translation of an artificial-intelligence-generated molecule from computational design to synthesis, preclinical validation, phase I testing and phase 2a clinical evaluation.

Advantages and Limitations

Generative artificial intelligence offers several potential advantages. It can explore chemical space rapidly, prioritize novel structures, support multi-property optimization and reduce unnecessary synthesis. It can also integrate heterogeneous information and assist researchers in selecting experiments with higher expected information value. These capabilities may reduce the number of design cycles required to identify a promising lead.

However, computational novelty does not guarantee biological novelty or therapeutic value. Models may reproduce biases present in training data, generate molecules that are unstable or difficult to synthesize, or overestimate activity because of dataset leakage or inadequate validation. Explainability is also important when predictions influence high-cost experimental decisions. Explainable artificial intelligence approaches aim to make model reasoning more interpretable, although interpretability remains an active research challenge [10].

Another important limitation is the difference between benchmark performance and real-world clinical success. A recent critical assessment of artificial intelligence in drug discovery concluded that clinically meaningful impact remains limited relative to the number of published computational advances and emphasized the need for better problem definition, stronger validation and greater focus on translational outcomes [21]. This does not diminish the value of the technology; rather, it indicates that future development should be driven by experimentally and clinically meaningful questions rather than by model novelty alone.

Table 3. Major advantages and challenges of generative artificial intelligence in drug discovery

Advantages	Challenges
Rapid exploration of chemical space	Limited or biased training data
Generation of novel molecular structures	Invalid or unstable generated molecules
Multi-property optimization	Conflicting objectives
Reduction in experimental search space	Synthetic accessibility problems
Prioritization of candidates	Prediction uncertainty
Support for medicinal chemistry decisions	Limited interpretability
ential reduction in design cycles	nslation from model to patient benefit
gration of experimental feedback	roducibility and regulatory requirements

Future Perspectives

The future of generative artificial intelligence in drug discovery is likely to depend less on producing increasingly complex models and more on integrating models with reliable experimental systems. Closed-loop design-make-test-learn workflows can continuously update computational models using experimentally confirmed results. Such systems may allow models to learn from both successful and unsuccessful experiments.

The integration of structural biology, phenotypic data, multi-omics information and high-content cellular measurements may improve biological context. Generative models may also become better at simultaneous optimization of potency, selectivity, solubility, metabolic stability, safety and synthetic feasibility. Human experts will remain important for defining meaningful objectives, recognizing biological inconsistencies and interpreting unexpected experimental results.

The emergence of artificial-intelligence-originated compounds in clinical testing provides evidence that the field is moving beyond purely theoretical demonstrations. Rentosertib is an important example because its development progressed from computational target discovery and molecular generation to preclinical validation and phase 2a patient testing [19,20]. Nevertheless, it remains essential to distinguish successful progression into clinical trials from regulatory approval or proven clinical superiority. As of the current evidence base, generative artificial intelligence should be regarded as a powerful discovery and prioritization strategy rather than a substitute for conventional drug development.

Conclusion

Generative artificial intelligence has introduced a new approach to de novo small-molecule drug discovery by enabling computational generation and optimization of molecular structures before laboratory synthesis. The technique can reduce the search burden associated with large chemical space, support multi-property optimization and help researchers prioritize experimentally promising candidates. Evidence from focused molecular library generation, deep generative design and more recent clinical-stage programs demonstrates meaningful progress. The development of rentosertib illustrates how an artificial-intelligence-driven program can progress from target discovery through preclinical studies and into patient trials. However, the ultimate value of the technique depends on data quality, experimental validation, synthetic feasibility, safety, clinical efficacy and regulatory confidence. The most effective future strategy is therefore not artificial intelligence alone, but a disciplined integration of generative models with medicinal chemistry, pharmacology, toxicology and clinical science.

References

Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. 2019;18(6):463–477.

Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov. Today. 2018;23(6):1241–1250.

Cerchia C, Lavecchia A. New avenues in artificial-intelligence-assisted drug discovery. Drug Discov. Today. 2023;28(4):103516.

Chen W, Liu X, Zhang S, Chen S. Artificial intelligence for drug discovery: resources, methods, and applications. Mol. Ther. Nucleic Acids. 2023;31:691–702.

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):513–530.

Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent. Sci. 2018;4(1):120–131.

Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. 2019;37:1038–1040.

Hasselgren C, Oprea TI. Artificial intelligence for drug discovery: are we there yet? Annu. Rev. Pharmacol. Toxicol. 2024;64:527–550.

Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent. Sci. 2019;5(9):1572–1583.

Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat. Mach. Intell. 2020;2:573–584.

Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589.

Thornton JM, Laskowski RA, Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. Nat. Med. 2021;27:1666–1669.

Tong AB, Burch JD, McKay D, et al. Could AlphaFold revolutionize chemical therapeutics? Nat. Struct. Mol. Biol. 2021;28:771–772.

Mullowney MW, Duncan KR, Elsayed SS, et al. Artificial intelligence for natural product drug discovery. Nat. Rev. Drug Discov. 2023;22(11):895–916.

Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat. Mater. 2019;18:435–441.

U.S. Food and Drug Administration. Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products. Discussion Paper. Center for Drug Evaluation and Research; 2023.

Mendez D, Gaulton A, Bento AP, et al. ChEMBL database in 2023: a drug discovery platform

spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52(D1)–D1192.

Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank knowledgebase for 2024. Nucleic Acids Res. 2024;52(D1)–D1275.

Ren F, Aliper A, Chen J, et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. Nat. Biotechnol. 2025;43(1):63–75.

Xu Z, Ren F, Wang P, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. Nat. Med. 2025;31(8):2602–2610.

Bender A, Thomas MC, Scannell JW, et al. Artificial intelligence in drug discovery — what it is, where we stand and the path forward. Nat. Rev. Drug Discov. 2026. doi:10.1038/s41573-026-01496-2.

Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: benchmarking models for de novo molecular design. J. Chem. Inf. Model. 2019;59(3):1096–1108.