Generative AI: Accelerating Novel Therapeutic Compound Design from *De Novo* Generation to Clinical Trials

Rasit Dinc

Rasit Dinc Digital Health & AI Research

Published: September 20, 2025 | AI Diagnostics

DOI: 10.5281/zenodo.17996547

Abstract

Generative AI: Accelerating Novel Therapeutic Compound Design from De Novo Generation to Clinical Trials The pharmaceutical industry faces a persistent ...

Generative AI: Accelerating Novel Therapeutic Compound Design from *De Novo* Generation to Clinical Trials

The pharmaceutical industry faces a persistent challenge: the escalating cost and time required to bring a new drug to market, often estimated at over $2.5 billion and spanning more than a decade [1]. The sheer size of the chemical space—the theoretical collection of all possible drug-like molecules—is astronomically large, making traditional high-throughput screening methods inefficient. In this context, **Generative Artificial Intelligence (GAI)** has emerged as a transformative technology, promising to revolutionize the early stages of drug discovery, particularly the design of novel therapeutic compounds.

The Paradigm Shift: From Screening to *De Novo* Generation

Historically, drug discovery has relied on screening vast libraries of existing compounds or fine-tuning known molecules. GAI introduces a paradigm shift by enabling ***de novo* molecular generation**—the creation of entirely new chemical structures with pre-defined, desirable properties [2]. This capability moves the process from a slow, iterative search to a rapid, goal-directed design.

GAI models are trained on massive datasets of known chemical compounds, learning the underlying "molecular grammar" and the complex relationships between a molecule's structure and its biological activity (ADMET properties: absorption, distribution, metabolism, excretion, and toxicity) [3]. By understanding these patterns, GAI can explore the chemical space more intelligently, generating novel candidates optimized for specific targets and characteristics.

Key Generative Models in Compound Design

The application of GAI in therapeutic compound design is primarily driven by several sophisticated deep learning architectures:

| Generative Model | Acronym | Mechanism in Drug Design | Key Advantage | | :--- | :--- | :--- | :--- | | **Variational Autoencoders** | VAEs | Encode molecules into a continuous, latent space, allowing for interpolation and sampling of novel, drug-like structures. | Stable training and a continuous latent space for fine-tuning properties. | | **Generative Adversarial Networks** | GANs | Use a generator and a discriminator network in a competitive process to produce highly realistic and novel molecular structures. | Ability to generate highly novel and diverse compounds. | | **Recurrent Neural Networks** | RNNs | Used in sequence-based generation (e.g., SMILES strings), building molecules one atom or bond at a time. | Effective for sequence-based molecular representations. | | **Large Language Models** | LLMs/CLMs | Chemical Language Models (CLMs) adapt transformer architectures to process molecular sequences, enabling complex conditional generation. | Leveraging advancements in NLP for molecular design, such as the development of tools like DrugGPT [4]. |

VAEs and GANs are currently the most popular GAI models for designing new therapeutic compounds [2]. VAEs, in particular, offer a continuous and organized latent space, which is crucial for controlling the properties of the generated molecules, such as solubility or target affinity. More recently, the adaptation of **Transformer** architectures, which power Large Language Models (LLMs), has led to the emergence of Chemical Language Models (CLMs), offering a more sophisticated approach to conditional generation [2].

Real-World Validation and Clinical Impact

The promise of GAI is rapidly moving from theoretical potential to clinical reality. A landmark example is the work by Insilico Medicine, which utilized its GAI platform, including the Chemistry42 module, to identify a novel drug candidate for Idiopathic Pulmonary Fibrosis (IPF). The GAI-driven process, from target identification to the nomination of a preclinical candidate, was completed in a fraction of the time and cost of traditional methods, and the molecule is now advancing through clinical trials [5].

This success story highlights the potential for GAI to dramatically compress the drug discovery timeline. By automating the design and optimization of lead compounds, GAI platforms can reduce the time from target to preclinical candidate from years to mere months.

Challenges and the Path Forward

Despite these breakthroughs, challenges remain. The primary hurdles include the **quality and quantity of training data**, the need for **experimental validation**, and the inherent **complexity of biological systems** [2].

1. **Data Scarcity and Bias:** GAI models require vast, high-quality datasets. For novel or "undruggable" targets, data can be scarce. Furthermore, if the training data is biased, the generated compounds may be unsafe or ineffective. 2. **Synthesizability:** A generated molecule must be chemically

feasible to synthesize in a lab. Researchers are actively working to integrate synthetic accessibility scores into the GAI models' objective functions to ensure the *de novo* designs are practical [6]. 3. **The "Black Box" Problem:** Deep learning models can be opaque, making it difficult to understand *why* a particular molecule was generated. Ongoing research is focused on developing more interpretable GAI models to build trust and guide medicinal chemists more effectively.

The future of therapeutic compound design lies in the synergistic collaboration between GAI, cheminformatics, and medicinal chemistry. By addressing the current limitations through techniques like **Transfer Learning** and **Reinforcement Learning** to refine molecular properties, GAI is poised to unlock previously inaccessible regions of the chemical space, leading to a new era of faster, more efficient, and more successful drug discovery.

**

References

[1] DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of Health Economics*, 47, 20-33.*

[2] Gangwal, A., et al. (2024). Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. Frontiers in Pharmacology*, 15. [https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2024.1331062/fu (https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2024.1331062/fu*

[3] Elton, D. C., Boukouvalas, Z., Fuge, M. D., & Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. Molecular Systems Design & Engineering*, 4(5), 828-849.*

[4] Li, Z., et al. (2023). DrugGPT: A large language model for drug discovery. arXiv preprint arXiv:2305.10689*.*

[5] Zhavoronkov, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology*, 37(9), 1038-1040.*

[6] Gao, W., & Coley, C. W. (2020). The synthesizability of molecules proposed by generative models. Journal of Chemical Information and Modeling*, 60(12), 5714-5723.