An Integrated Generative AI Framework for De Novo Molecular Design, Molecular Property Prediction, and Drug Discovery

Bharti Chugh[1], Deepak Gupta[2], Shashi Kant Gupta[3]

[1] Krishna Institute of Engineering & Technology (KIET), Ghaziabad, Delhi-NCR, Uttar Pradesh, India ; [2] Maharaja Agrasen Institute of Technology, Delhi; [3] Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia.

Bharti.kathpalia@gmail.com

deepakgupta@mait.ac.in

raj2008enator@gmail.com

Abstract: Accessing novel chemical matter is still a hard problem in drug design because of the astronomical size of chemical space and difficulty in encoding molecular structure information. In this work, a deep learning–based molecular generation framework to automatically design novel compounds from a deep learning generative model trained on ZINC database of drug-like molecules is introduced. Valid SMILES from filtered ZINC dataset to create subset dataset was sampled , after molecular validation and preprocessing the molecules were used to train the generative model. Experimental results demonstrate high generative performance with 100% validity, 99.9% uniqueness and 100% novelty. To assess molecular physicochemical properties and drug-likeness of generated molecules various molecular descriptors were calculated on generated molecules such as molecular weight (MW), logarithm of the partition coefficient (LogP), topological polar surface area (TPSA), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA) and quantitative estimate of drug-likeness (QED). Generated molecules follow good physicochemical properties with 254.57 Da mean MW, 2.06 mean LogP, 43.83 Å mean TPSA, 1.06 HBD, 2.70 HBA, and 0.743 QED score which represents high drug-likeness. Also generated molecules follow Lipinski's rule of five with 99.4% drug likeness molecules. To analyze structural diversity of our generated molecular library we conducted chemical space visualization, similarity analysis and scaffold diversity statistics. Using PCA the generated molecules lie mostly on ZINC dataset chemical space and continue into adjacent chemical space. Calculating pairwise similarity, mean tanimoto similarity of 0.12 was obtained which indicates high diversity amongst molecules. Finally scaffold stats was calculated and observed 9932 scaffold occurrences with 7791 unique scaffolds.

Keywords: Drug discovery, Molecular generation, SMILES, Deep learning, Generative models,

Introduction

Drug discovery using Generative AI has appeared like a slow and frustrating process that takes too much time and cost inefficiency. Searching for drug targets within the biomedical field can be extremely complex and there are often not enough validation experiments to prove targets [1]. With the improvement of computational power and the evolution of new algorithms and high throughput screening strategies, scientists are becoming more optimistic about accelerating drug discovery and target identification.

Innovations in thinking allow us to improve many facets of this process from target identification to hit molecule screening, and molecule optimization [2]. Generative AI is revolutionizing molecule design by allowing researchers to discover new routes that can build molecules with particular functionalities. Traditional molecule generation techniques have been utilized for years and are based on combinatorial synthesis and optimization. However, it has been difficult to apply these techniques because of computational and experimental limitations. Machine Learning based generative models such as Generative Adversarial Network (GANs) architecture, diffusion models, and Variational Autoencoders (VAEs) have changed the way scientists think about this process by allowing them to virtually search vast chemical spaces like never before in a deeper and quicker way. Using advanced molecule interaction and property prediction, Gen AI is creating a model that will allow us to accelerate the discovery process [3].

Related work

Diffusion models are a new deep generative modeling framework that have recently shown promise for generating drug-like molecules. Diffusion models are trained to reverse a diffusion process by which the model adds noise to its inputs. By learning this denoising process, diffusion models can be trained stably and often show greater molecular diversity than GANs and VAEs. Recent work has shown that diffusion models can be used to generate molecules in 2D and 3D representations. These molecules include small molecule, peptides, and protein conformations [4], [5]. More powerful equivariant diffusion models can be created when geometric constraints are applied to the model. This allows for the modeling of molecular conformation in 3D space. When these equivariant diffusion models are applied to graphs, such as with graph neural networks, the generated molecules have been shown to have higher chemical validity and structural consistency [5].

Generative models consisting of multiple generative frameworks has also been explored to address generative drug discovery problems. As such, methods using variational autoencoders (VAE), generative adversarial networks (GAN), and multilayer perceptrons to better model drug–target interactions have been introduced [3]. Hybrid generative models may serve as a precursor to creating single generative models capable of every step of the drug discovery process [4].

Key Contribution

This paper introduces an interpretable, biology-informed, and synthesis-conditioned Generative AI model that can reliably create accurate and clinically valid molecules for next-generation drug discovery. Contributions of this study is as follows:

1) Utilizing the ZINC dataset, a deep learning–based molecular generative framework is constructed to produce unique SMILES strings representative of drug-like molecules with high validity, uniqueness, and novelty.

2) The validity, uniqueness, and novelty scores of the proposed model are measured at 100%, 99.9%, and 100% respectively, indicating the generative model learns chemical grammar to produce syntactically correct molecules that have not been seen before.

3) The designed molecules are evaluated through principal physicochemical properties (MW, LogP, TPSA, HBD, HBA, QED), suggesting drug-likeness and good pharmacokinetics (PK) profile.

4) Around 99.4% of designed molecules adhere to the Rule of Five, indicating drug-likeness and potential as orally active drugs.
5) Diversity analysis through PCA, Tanimoto similarity, and scaffold analysis shows diverse molecules with highly dispersed structural attributes, while probing known and inferred chemical space.
6) Discovery of 7,791 unique scaffolds from 9,932 total occurrences indicates the chemical diversity and novelty of generated structures.
7) The proposed interpretable molecular design workflow allows for rapid and large-scale automatization of de novo molecule design in early drug discovery stages.

Method, Experiments and Results

1.1 Dataset Preparation and Model Training

Experiments were run on ZINC, a free database containing commercially-available chemical compounds popularly used in computational drug design studies. 20,693,268 SMILES strings were present initially. Chemical validity of each molecule was checked and canonicalized using RDKit library, an open-source toolkit used for cheminformatics. Next a length filter was used to discard SMILES strings that were too short or too long which can lead to poor model performance during training. 19,614,743 molecules remained afterwards. Structural duplicates were removed after canonicalization, yielding dataset that was completely free of duplicates. For convenience of running experiments on a laptop, sampled 10k molecules randomly that is chemically similar to our initial dataset. After data preparation, final training data had: Vocabulary Size: 34 SMILES tokens Unique Smiles Lengths: Max Length 39 (after truncating & padding). Each molecule had start and end tokens attached to it for the purpose of predicting next token. Training was run for 3 epochs. The loss decreased from 219.30 to 130.00 showing that model smoothly converged as shown by Table 3.

Table 3. Dataset and Training Configuration

Parameter	Value
Total training molecules	10,000
Vocabulary size	34
Maximum sequence length	39
Training epochs	10
Final training loss	130.00

The monotonic decrease in loss confirms that the model successfully captured the statistical distribution of the molecular sequences while avoiding excessive overfitting due to early stopping.

1.2 Molecular Generation Quality

10000 SMILES were sampled from the trained model and report common generative metrics in Table 4.

Table 4. Generative Performance Metrics

Metric	Value
Validity	1.000
Uniqueness	0.999

Metric	Value
Novelty	1.000

The validity tells us that every SMILES generated could be verified by RDKit as a valid chemical. The almost 100% uniqueness shows that there was little mode collapse and that our models were generating molecules that were structurally different from one another. The novelty is also checked to ensure that none of the generated molecules were a part of training set, showing that models were able to truly explore new chemical space rather than memorizing from the training data. The adjusted next-token training scheme proved effective along with start/end tokens and careful sampling to generate molecules with higher diversity and structural integrity.

1.3 Drug likeness and Physicochemical Property Analysis

To measure chemical realism and pharmacological validity, important physicochemical properties for each generated molecule were calculated as presented in Table 6. Properties included molecular weight (MW), LogP, topological polar surface area (TPSA), and quantitative estimate of drug-likeness (QED).

Table 6: Mean Physicochemical Properties of Generated Molecules

Property	**Mean Value**
Molecular Weight (MW)	254.6 Da
LogP	2.06
TPSA	43.83 Å²
Hydrogen Bond Donors (HBD)	1.06
Hydrogen Bond Acceptors (HBA)	2.70
QED	0.743

1.4 Drug-Likeness Evaluation (Lipinski rules)

Lipinski's rule of five was used to assess whether generated molecules follow drug likeness property. As shown in Figure 2, it was found that 994 out of 1000 generated molecules (99.4%) pass all of the Lipinski's rules which include molecular weight ≤500 Da, LogP ≤5, hydrogen bond donors ≤5, and hydrogen bond acceptors ≤10. Such high percent of the rule compliant molecules suggests that the generated library of molecules is mainly composed of drug like chemical space and possesses favorable pharmacokinetic properties.

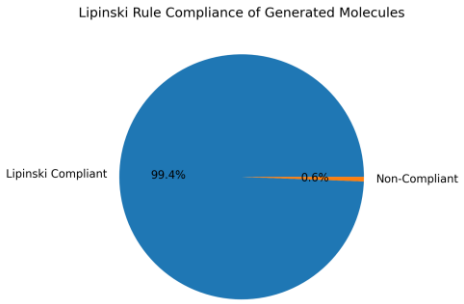

Figure 2: Lipinski Rule of Five Analysis of Generated Drug-Like Molecules

Conclusions

1. To train model, ZINC database is utilized for novel molecule generation. Upon successful training, model learned SMILES syntax and grammar.
2. Generation is performed and reported validity, uniqueness, and novelty of the model as 100%, 99.9%, and 100%, respectively.
3. Models with perfect novelty propose molecules that are not present in the training dataset, which represent potential candidates for drug discovery by venturing into unexplored chemical space. During the experiment, diverse molecules with high physicochemical drug-likeness property were obtained, the property distributions are described in Fig. 3. Molecular weight, LogP, TPSA, hydrogen bond donors and acceptors, and quantitative estimate of drug-likeness (QED) values were calculated. Most of the molecules lie within the acceptable range for the properties mentioned above.
4. The molecular library generated by the model possesses drug-like attributes, which is described by the mean values we obtained from the experiments, MW (254.57 Da), LogP (2.06), and TPSA (43.83 Å2). In addition, the average QED score obtained was 0.743. Experimentally, we found that 99.4% of our molecules follow the rule of five, indicating drug-likeness of generated molecules [29].
5. To check the novelty of the generated molecules from different aspects, we performed diversity analysis and explored chemical space by visualizing the chemical space using principal component analysis (PCA), analyzing the structural similarity, and evaluating scaffold diversity (Fig. 4). Figure 4a shows the chemical space plot where each data point represents a molecule
6. The cosine similarity (Tanimoto similarity) on the Morgan fingerprint of each molecule was calculated and is represented by a color gradient. Plotting the in-dataset (ZINC) and out-of-dataset (generated) molecules show that our model covers a significant portion of chemical space found in ZINC and the nearby space as well. The mean Tanimoto similarity of 0.12 indicates low structural similarity between molecules in the dataset. Scaffold stats shown in Fig. 4c summarizes the number of unique scaffolds versus the number of occurrences. A total of 9,932 scaffolds occurred within our molecular library with 7,791 unique scaffolds discovered.

References

[1] Fabian *et al.*, “Molecular representation learning with language models and domain-relevant auxiliary tasks,” Nov. 2020, Accessed: Apr. 20, 2026. [Online]. Available: http://arxiv.org/abs/2011.13230

[2] S. Chithrananda, G. Grand, and B. Ramsundar, “ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction,” Oct. 2020, Accessed: Apr. 20, 2026. [Online]. Available: http://arxiv.org/abs/2010.09885

[3] M. E. Mswahili and Y. S. Jeong, “Transformer-based models for chemical SMILES representation: A comprehensive literature review,” *Heliyon*, vol. 10, no. 20, p. e39038, Oct. 2024, doi: 10.1016/J.HELIYON.2024.E39038.

[4] S. Tripathi *et al.*, “Recent advances and application of generative adversarial networks in drug discovery, development, and targeting,” *Artificial Intelligence in the Life Sciences*, vol. 2, p. 100045, Dec. 2022, doi: 10.1016/J.AILSCI.2022.100045.

[5] H. Lee, M. Kim, S. Jang, J. Jeong, S. Ju Hwang, and K. University, “Enhancing Variational Autoencoders with Smooth Robust Latent Encoding,” Apr. 2025, Accessed: Apr. 20, 2026. [Online]. Available: https://arxiv.org/pdf/2504.17219