ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

GAMS: Graph-Based Autoregressive Molecular Generation System for Drug Discovery

Jindong Sha

Sage Hill SchooI UNITED STATES OF AMERICA, Newport, California, 92657, United States of America

Abstract: *Generating novel molecular structures with pharmacological activity remains a fundamental challenge in drug discovery. Traditional drug development approaches, heavily reliant on manual labor, are characterized by time-consuming processes, substantial costs, and limited exploration efficiency. To address these limitations, we propose a Graph-based Molecular Generation System (GAMS) that integrates decoder architecture with graph structures to effectively capture molecular structural features, physicochemical properties, and pharmacological characteristics, enabling rapid generation and design of drug-like molecules. Our innovation lies in the development of a dynamic property-guided decoding strategy that periodically incorporates drug property predictions during the generation process, enhancing the drug-likeness of generated molecules. Experimental results demonstrate that GAMS outperforms some existing methods in terms of quantitative drug-likeness scores (QED) and the proportion of drug-like molecules, validating its effectiveness in molecular design and optimization.*

Keywords: *Drug Discovery, Molecular Generation, Autoregressive Model, Deep Learning*

1. Introduction

Drug development is an extremely complex, time-consuming, costly, and high-risk process. According to a comprehensive study by DiMasi et al., the average cost of developing a new drug from laboratory to market amounts to US$2.87 billion (2013 dollars), with this cost increasing at an annual rate of 8.5%, significantly outpacing general inflation. More concerningly, the clinical success rate is merely 11.83%, indicating that nearly nine out of ten drug candidates entering clinical phases ultimately fail to obtain approval [1].

Berdigaliyev and Aljofan demonstrate that bringing a new drug from concept to market typically requires 10-15 years of development. This lengthy process encompasses target identification, lead compound screening, drug optimization, preclinical studies, multiple phases of clinical trials, and regulatory review, with each stage subject to rigorous scientific and regulatory requirements. The Center for Drug Evaluation and Research (CDER) of the U.S. Food and Drug Administration (FDA) ensures that every new drug undergoes thorough safety and efficacy evaluation before market approval [2].

The strategies and technologies for drug discovery have evolved over generations.According to [3], the advent of high throughput screening (HTS) has significantly transformed the drug discovery landscape, allowing for the rapid screening of large compound libraries. Despite the widespread adoption of new technologies such as ultra-high-throughput drug screening and computer-aided drug design, the overall cost and timeline of drug development remain largely unchanged.

In response to these challenges, artificial intelligence (AI), particularly deep learning (DL), has emerged as a promising solution in drug development. DL, as a subset of AI that employs multiple layers of nonlinear processing units, has demonstrated exceptional capabilities in handling vast amounts of biomedical data. [4] Specifically, deep learning (DL) algorithms have been implemented in various drug discovery processes, including peptide synthesis, structure-based and ligand-based virtual screening, toxicity prediction, drug monitoring and release, pharmacophore modeling, quantitative structure–activity relationship, drug repositioning, polypharmacology, and physiochemical activity.[5]

Current methods face significant challenges in generating molecules with desirable pharmacological properties, often resulting in compounds of insufficient quality that fail to meet the stringent requirements of drug development. Our approach introduces graph-structured directed

ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

optimization and attribute-guided decoding, enabling the more effective generation of high-quality pharmacological molecules. This provides a novel perspective for drug development.

2. Related Work

The work[6] has extensively investigated the performance of RNN-based molecular generation models across various configurations. Researchers benchmarked different SMILES representations (canonical, randomized, and DeepSMILES), dataset sizes (1M, 10K, and 1K molecules), and model architectures (LSTM and GRU cells) while exploring different hyperparameter settings such as dropout rates and batch sizes. They introduced novel metrics to evaluate model generalization by assessing the uniformity, closedness, and completeness of the generated chemical space. Their results demonstrated that LSTM models trained on 1M randomized SMILES sequences achieved superior performance in molecular generation tasks.

Recent research[7] proposed a novel SMILES-based molecular generative architecture for scaffold-based molecule generation. The approach combines a scaffold generator and a decorator model with an innovative molecular slicing algorithm, achieving 0.454-0.989 predicted active molecules on the DRD2 modulator dataset. This work opens new possibilities for scaffold-based drug design.

This paper [8] presents a novel deep learning approach to enhance the drug discovery process. The researchers developed a BiLSTM Attention Network (BAN) incorporating a multi-step attention mechanism to extract molecular features from SMILES strings. To address the data scarcity challenge, they implemented SMILES enumeration for data augmentation and utilized the same enumeration technique during the prediction phase to improve accuracy. The method demonstrated superior performance across 11 ADMET-related tasks (including absorption, distribution, metabolism, excretion, and toxicity) compared to state-of-the-art approaches, indicating significant potential for practical applications in pharmaceutical research.

[9] Proposes a novel molecular representation method based on multiple SMILES augmentation for improving molecular property prediction tasks. The researchers generate multiple SMILES sequences for each molecule as an automated data augmentation strategy, effectively alleviating the overfitting problem caused by small sample sizes in molecular property prediction datasets. The method integrates stacked CNN and RNN network architectures, eliminating the need for descriptor engineering and expert experience while enabling more comprehensive learning of SMILES grammatical features. Experimental results demonstrate that the multiple SMILES-based augmentation approach achieves superior performance across various molecular property prediction tasks, providing new insights for deep learning applications in drug discovery.

3. Method

3.1 Foundational Framework

The performance of the Decoder model has been widely validated in the field of Natural Language Processing (NLP), with nearly all large language models (LLMs) based on the Decoder architecture. Considering that SMILES (Simplified Molecular Input Line Entry System) represents the two-dimensional structure of molecules through atomic symbols, bond types (such as single bonds, double bonds, etc.), as well as branching and cyclic structures, the string data format is naturally suitable for Decoder input. This allows us to leverage mature techniques and methodologies from the NLP domain to pre-train SMILES molecular representations, thereby enhancing their chemical generation capabilities.

Let the SMILES string be represented as a discrete symbol sequence $S=(s_1,s_2,\ldots,s_T)$, where each symbol $s_t \in \Sigma$ belongs to a finite character set (including atoms, bond types, branching symbols, etc.). In this framework, the model employs an autoregressive approach to model the joint probability distribution:

$$P_\theta(S)=\prod_{t=1}^{T} P_\theta(s_t \mid s_{<t})$$

Where θ represents the model parameters, and the conditional probability $P_\theta(s_t \mid s_{<t})$ is computed

ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

by the Decoder. In this manner, the molecular generation task is transformed into a character-by-character prediction task, where the model relies on previously generated symbols to capture the sequential and dependency relationships of the molecular structure.

To effectively model long-range dependencies in context, the model utilizes an attention mechanism. This mechanism enables the model to consider the entire context of the sequence when generating each symbol, thereby improving the accuracy and rationality of the generation. Specifically, the topological structure of the SMILES string is modeled in conjunction with the attention mechanism using positional encoding $E_{pos} \in \mathrm{R}^{T_{max} \times d}$.

We define the contextual representation of the symbol s_i as:

$$\mathrm{h}_i(L)=\mathrm{Transformer}(E_{token}(s_i)+E_{pos}(i))$$

Where L denotes the total number of layers. The high-order representation $\mathrm{h}_i(L)$ encodes the topological features of the molecular substructure. Through the projection matrix $W_o \in \mathrm{R}^{d \times |\Sigma|}$, it is mapped to the vocabulary space, and the Softmax function is applied to generate the probability distribution for the next symbol:

$$P_\theta(s_t \mid s_{<t})=\mathrm{Softmax}(W_o^{\mathrm{T}}\mathrm{h}_t(L))$$

This process transforms the semantic information of the hidden state into likelihood scores for candidate symbols, thereby enabling autoregressive generation.

During the training process, we adopt the cross-entropy loss function to evaluate the model's performance. Specifically, the loss function is defined as:

$$\mathrm{L}(\theta)=-\frac{1}{T}\sum_{t=1}^{T}\log P_\theta(s_t \mid s_{<t})$$

This loss function guides the model to learn more accurate symbol generation strategies by minimizing the discrepancy between predicted symbols and true symbols. Through this pre-training approach, the model can learn the correct representations of SMILES molecules from a vast data space, thereby enhancing its performance in chemical generation tasks.

3.2 Dynamic Guided Decoding

To further enhance the Decoder model's capability in generating high-quality SMILES molecules, we propose an innovative dynamic guided decoding method based on property prediction. Unlike traditional autoregressive decoding, this approach achieves directional control over the molecular generation process by dynamically incorporating property prediction during decoding.

Specifically, we designed an adaptive strategy at the decoding stage that directionally guides the decoding path through the integration of additional property predictions, thereby better ensuring the drug-likeness of the generated molecules. To achieve this objective, we first trained a pharmaceutical property discriminator model to evaluate the pharmacological properties of SMILES fragments:

$$f(\theta):\phi(s)\rightarrow[0,1]$$

Where $\phi(s) \in \mathrm{R}^d$ represents the high-dimensional feature representation of molecular fragments, capturing key chemical features and pharmacological properties.

To implement dynamic guidance, we innovatively introduced a periodic property prediction mechanism, defining the guidance trigger time set as:

$$t \in T_K=\{k \cdot K \mid k \in \mathrm{N}^*\}$$

This periodic triggering mechanism ensures the stability of property prediction and the effectiveness of long sequences while avoiding excessive intervention in the original generation process. At time step t-1, we obtain the token sequence $\{1{:}x_{t-1}\} \in \mathrm{R}^{t-1}$ and generate the corresponding SMILES fragment $s(\{1{:}x_{t-1}\})$. A key innovation lies in our design of an evaluation mechanism that, for each candidate token x_i, employs the scoring function:

$$I(t,i)=f_\theta(\phi(s(\{1{:}x_{t-1}\} \oplus x_i)))$$

ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

This scoring function considers the current state and evaluates the pharmacological properties of potential molecular fragments formed after adding new tokens, implementing prospective guidance for the generation process. Based on this, we propose an adaptive logits adjustment strategy:

$$\text{logits}(t,i)=\text{ori}(t,i)+\alpha I(t,i)$$

Where α is a dynamic regulation factor that balances the relationship between the original generation distribution and property guidance. The final conditional probability distribution is calculated through softmax normalization:

$$P(x_t|\phi(1:t\text{-}1))=\frac{\exp(\text{logits}(t,i))}{\sum_j \exp(\text{logits}(t,j))}$$

As shown in Figure 1, this decoding strategy dynamically integrates property prediction at certain decision points, evaluating candidate token sequences through the discriminator model to obtain potential pharmacological scores and accordingly directing the generation probabilities. This approach not only enhances the quality and drug-likeness of the generated molecules but also maintains the flexibility and diversity of the generation process.

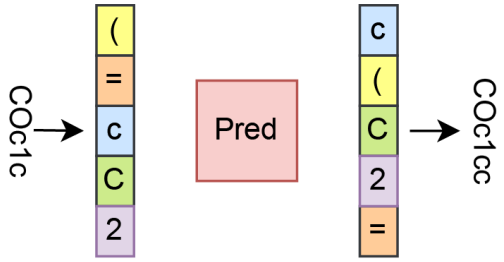

Figure 1: Property guidance

3.3 Graph Structure

Based on the aforementioned methods, we have trained a Decoder model guided by attributes. In the process of drug molecule generation, it is often necessary to generate a large number of candidate molecules for preliminary screening, and these molecules often have certain intrinsic relationships. Therefore, this study chooses to use a graph structure to represent the molecular optimization space, as the graph structure can effectively store the relationships between molecules, such as similarity, structural changes, derivation paths, and the physicochemical properties of the molecules. This representation helps to expand and classify families of molecules with similar properties.

In the molecular optimization process, we need to ensure that the molecules maintain appropriate similarity. An accurate similarity measurement mechanism is crucial for guiding the search direction and ensuring optimization quality. Traditional methods often focus solely on either structural similarity or property similarity, which may lead to loss of activity or insufficient improvement of properties during optimization. Therefore, we propose a new dual similarity measurement method that unifies the dual objectives of structural conservativeness and property optimization in molecular similarity calculations. This method can be expressed as:

$$\text{Sim}(v_i,v_j)=\alpha\cdot T_{\text{animoto}}(f_{p_i},f_{p_j})+\beta\cdot\cos(a_i,a_j)$$

Here, we use Morgan Fingerprint (f_{p_i},f_{p_j}) as the basis for molecular structural representation. The Morgan Fingerprint captures the local chemical environment information of atoms through an iterative updating method based on circular environments. By adjusting the radius parameter (r=1,2,3), we can achieve hierarchical representation of structural features at different spatial scales, retaining key local structural patterns while incorporating broader molecular backbone information. We choose to generate a 2048-bit binary fingerprint vector, which strikes a good balance between information content and computational efficiency.

In terms of structural similarity measurement, we selected the Tanimoto coefficient as the evaluation metric. The Tanimoto coefficient is one of the most widely used similarity measures in the field of cheminformatics, and its mathematical expression is:

ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

$$T_{\text{animoto}}(A,B)=\frac{|A\cup B|}{|A\cap B|}$$

The Tanimoto coefficient effectively captures key structural information such as the backbone structure of the molecule, substituent distribution, and local chemical environment by calculating the intersection-over-union ratio of Morgan fingerprints, providing reliable guidance for maintaining active groups and key pharmacophores.

In terms of property similarity measurement, we use cosine similarity to quantify the property differences between the original and evolved molecules:

$$\cos(a_i,a_j)=\frac{//\,a_i\,//\cdot\|a_j\,//}{(a_i\cdot a_j)}$$

The property vector comparison mechanism based on cosine similarity comprehensively considers multiple physicochemical parameters such as molecular weight, LogP, and polar surface area, assessing the degree of property change through their projection relationships in high-dimensional attribute space, thereby ensuring the rationality of the optimization direction. By adjusting the weight coefficients to balance structural properties and target attributes, we encourage structural innovation while expecting it to maintain similar properties.

To achieve efficient exploration and directed optimization, refer to Figure 2, we designed a node expansion strategy. This strategy fully utilizes the aforementioned dual measurement mechanism, achieving directed control of the optimization path by simultaneously considering the conservativeness of molecular structure and the optimization direction of properties during the expansion process.

The primary step in the node expansion process is to prioritize and select candidate nodes based on a comprehensive scoring mechanism. Specifically, we use a node scoring function:

$$\text{SelectScore}(v_i)=\lambda_1\cdot\text{Sim}(v_i,v_{\text{ori}})+\lambda_2\cdot\Delta\text{Property}(v_i)$$

Where $\text{Sim}(v_i,v_{\text{target}})$ employs the aforementioned dual similarity measurement to evaluate the similarity between the candidate node and the target node. $\Delta\text{Property}(v_i)$ measures the improvement of node v_i in the target property dimension, providing guidance for the expansion direction by assessing the property enhancement of the current node relative to its parent node and the dual similarity.

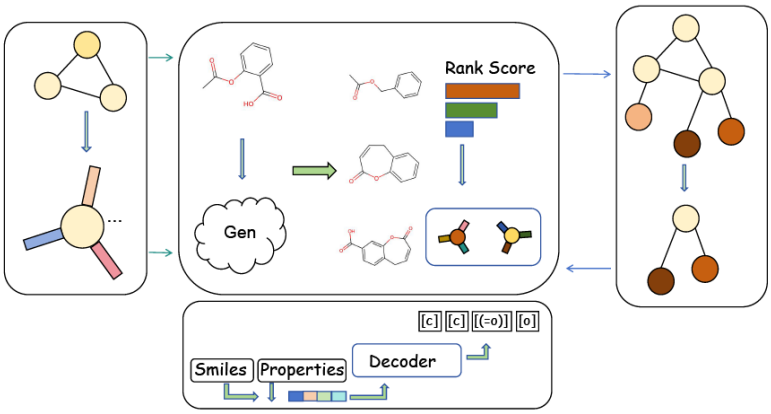

Figure 2: GAMS

4. Experiment

4.1 Dataset

We collect a comprehensive dataset for molecular property prediction by collecting and processing compound data from the pubchem database. The dataset comprises over 60,000 molecular entries, with each entry represented by its SMILES notation as input features and its corresponding QED as the

ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

target variable.

The raw molecular data was retrieved from the pubchem database, followed by rigorous data cleaning processes to ensure data quality and reliability. For each molecule, we computed its QED score to serve as our prediction target. The QED score ranges from 0 to 1, providing a quantitative measure of a molecule's likelihood to exhibit drug-like properties.

To ensure model robustness and generalizability, we meticulously engineered the dataset distribution to maintain balanced QED values across the entire possible range (0-1). We intentionally avoided concentrated distributions that could potentially introduce bias. This strategic sampling approach mitigates bias in the model training process and ensures that the model can effectively learn to predict drug-likeness across diverse molecular structures with varying QED scores.

4.2 Results

This study employs a Transformer-based deep regression model for quantitative drug-likeness (QED) prediction using a dataset containing 60,000 molecular structures. As shown in the results, the model demonstrates excellent predictive performance and robust generalization capabilities: the training and test loss functions converge to similarly low levels, indicating effective learning of data features while maintaining good generalization ability.

Performance evaluation shows that the mean absolute error (mae), which reflects the absolute deviation between predicted and experimental values, is calculated as:

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$$

Where y_i is the experimental value, $\hat{y}_i$ is the predicted value, and n is the number of samples. The MAE of 0.04 indicates that the average deviation between model predictions and experimental values is controlled within 4%, meeting the practical requirements for computational prediction accuracy in drug discovery.

The Coefficient of Determination (R^2), which measures the proportion of variance in the target variable explained by the model, is defined as:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Where $\bar{y}$ represents the mean of experimental values. The R^2 value of 0.93 demonstrates that the model explains 93% of the variance in the target variable, indicating high statistical consistency between predictions and true values , as is evident in Figure 3.

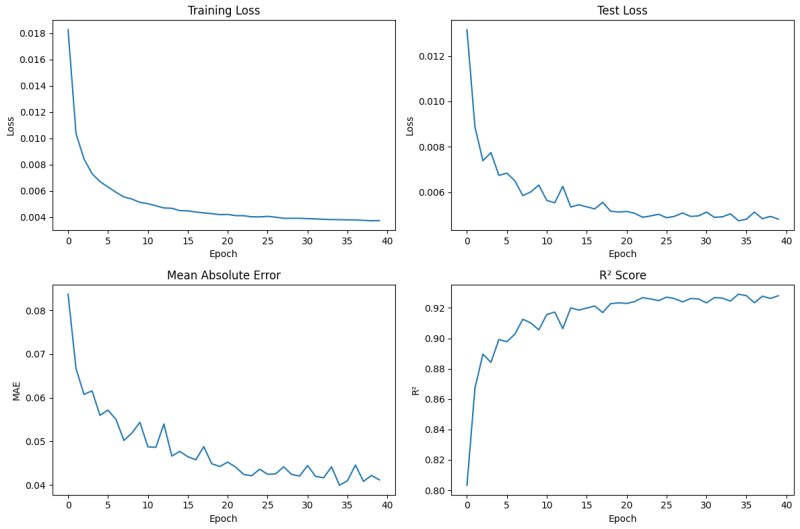

Figure 3: Metrics

In our attribute-guided decoding strategy, we implemented an intervention interval parameter K=10,

ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

wherein the candidate token sequences undergo attribute-guided selection every 10 steps during the generation process. This design enables directed guidance toward high-quality search paths while preventing excessive intervention in the fundamental generation process. We conducted a comprehensive evaluation by generating thousands of SMILES molecular samples using both a standard decoder model and our improved algorithm with attribute guidance, systematically assessing their performance through mean QED calculations and molecular validity verification.

	Decoder	Guide
QED	0.55	0.62
Validity	0.95	0.91

Figure 4: Comparison between decoder and property guidance

As illustrated in Figure 4, the attribute-guided decoding strategy demonstrated substantial improvement in the mean QED values of generated molecules, showing significant advantages over the baseline decoder. These results validate the effectiveness of our attribute-guided approach in enhancing the model's capability to generate high-quality drug-like molecules. Notably, we observed a slight decrease in molecular validity metrics, which can be attributed to the necessary perturbation of the original generative distribution introduced by the selective preferences during the attribute-guided process. Specifically, while this method effectively steers the model toward chemical space with enhanced drug-likeness properties, the introduced forced selection mechanism partially influences the original generation process. However, considering the substantial improvement in drug-likeness properties, this minor reduction in validity represents an acceptable trade-off that does not significantly impact the overall effectiveness of our method.

	Guide	Graph
QED	0.62	0.67
Drug-like molecules	0.94	0.97
Diversity	0.95	0.92

Figure 5: Property guidance and Graph

As illustrated in Figure 5, GAMS model compared to the attribute-guided decoder, achieving higher mean QED values and an increased proportion of drug-like molecules. This enhanced performance can be attributed to GAMS's sophisticated node expansion strategy, which prioritizes the exploration of nodes exhibiting favorable property improvements during molecular construction. This approach enables more efficient directed search within the chemical space, thereby optimizing target properties more effectively. The observed decrease in molecular diversity metrics can be explained through the same mechanism: when the model preferentially selects similar high-quality nodes for expansion, it may result in a certain degree of redundancy in the generated molecular expressions.

As demonstrated in Figure 6, our proposed methodology exhibits superior performance compared to existing models in terms of both mean QED values and the proportion of drug-like molecules.

ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

	Pocket2Mol	TargetDiff	Ours
QED	0.56	0.59	0.67
SAS	3.10	3.92	3.13
Drug-like molecules	0.95	0.86	0.97
Validity	1.0	0.98	0.99
Diversity	0.95	0.96	0.92

Figure 6: Models Compare

In this statistical analysis of the generation process, we evaluated the distribution characteristics of four key molecular properties. The Quantitative Estimate of Drug-likeness (QED) shows a distribution range of 0.60-0.85, with the majority concentrated in the 0.60-0.70 range, indicating that the generated compounds possess favorable drug-like properties. The Synthetic Accessibility (SA) Score ranges from 1.5 to 4.0, exhibiting a notable right-skewed distribution with predominantly low scores, suggesting moderate synthetic feasibility for the generated molecules. The molecular weight distribution spans from 180 to 260 Da, fully complying with Lipinski's Rule of Five requirement of molecular weight less than 500 Da, which is conducive to drug absorption and transport. The octanol-water partition coefficient (LogP) distribution falls within 0.0-2.0, lying within the recommended range (-0.4 to 5.6) and showing relatively uniform distribution characteristics. This indicates that the generated compounds possess appropriate lipophilicity, favorable for their distribution and transport within biological systems, which is visually confirmed in Figure 7. In conclusion, the compounds generated in this process demonstrate promising characteristics in terms of medicinal chemistry-related properties, suggesting potential value for drug development.

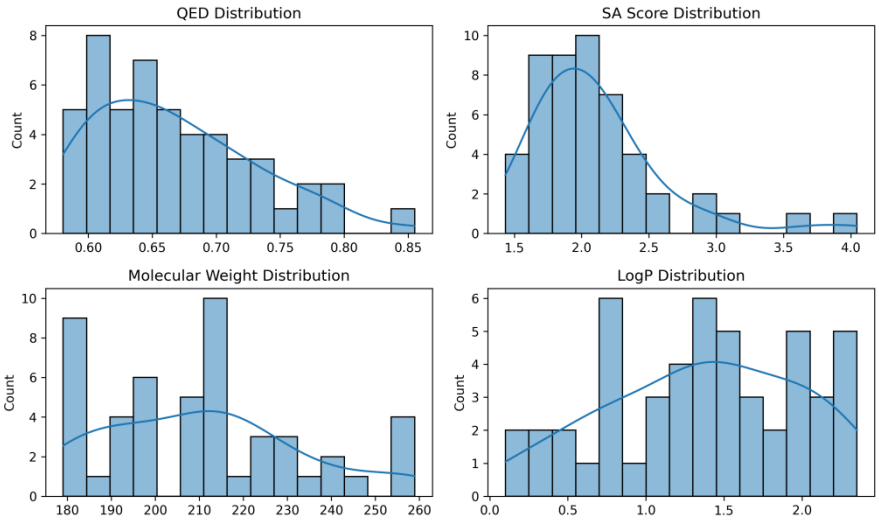

Figure 7: Properties

5. Conclusion

This study presents GAMS (Graph-based Autoregressive Molecular Generation System), an innovative framework that addresses critical challenges in AI-driven drug discovery by integrating decoder architectures with graph-based optimization. Our key contributions lie in two aspects: (1) A dynamic property-guided decoding strategy that periodically incorporates pharmacological predictions during generation, enabling directional control over molecular design; (2) A graph-structured

ISSN 2618-1584 Vol. 7, Issue 6: 32-40, DOI: 10.25236/FMSR.2025.070604

optimization framework that maintains structural conservativeness while pursuing property enhancement through dual similarity metrics; A node expansion mechanism that balances structural diversity and property optimization through adaptive scoring functions.

References

[1] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of r&d costs," Journal of health economics, vol. 47, pp. 20–33, 2016.
[2] N. Berdigaliyev and M. Aljofan, "An overview of drug discovery and development," Future medicinal chemistry, vol. 12, no. 10, pp. 939–947, 2020.
[3] D. A. Pereira and J. A. Williams, "Origin and evolution of high throughput screening," British journal of pharmacology, vol. 152, no. 1, pp. 53–61, 2007.
[4] A. Lavecchia, "Deep learning in drug discovery: Opportunities, challenges and future prospects," Drug discovery today, vol. 24, no. 10, pp. 2017–2032, 2019.
[5] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: Machine intelligence approach for drug discovery," Molecular diversity, vol. 25, pp. 1315–1360, 2021.
[6] J. Arús-Pous et al., "Randomized SMILES strings improve the quality of molecular generative models," Journal of cheminformatics, vol. 11, pp. 1–13, 2019.
[7] J. Arús-Pous et al., "SMILES-based deep generative scaffold decorator for de-novo drug design," Journal of cheminformatics, vol. 12, pp. 1–18, 2020.
[8] C.-K. Wu, X.-C. Zhang, Z.-J. Yang, A.-P. Lu, T.-J. Hou, and D.-S. Cao, "Learning to SMILES: BAN-based strategies to improve latent representation learning from molecules," Briefings in bioinformatics, vol. 22, no. 6, p. bbab327, 2021.
[9] B. Winter, C. Winter, J. Schilling, and A. Bardow, "A smile is all you need: Predicting limiting activity coefficients from SMILES with natural language processing," Digital Discovery, vol. 1, no. 6, pp. 859–869, 2022.