An-Najah University Journal for Research – A
Natural Sciences

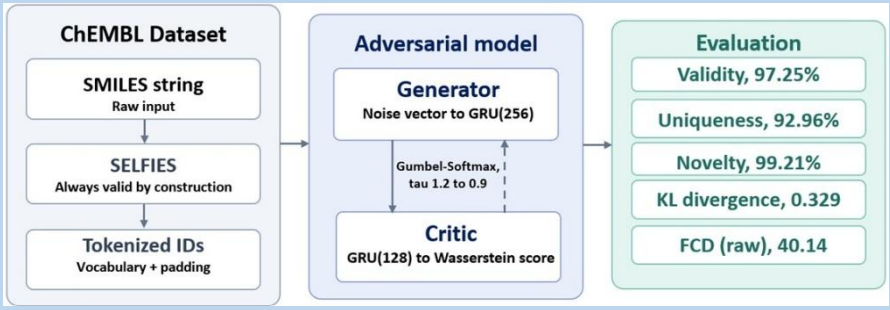
Discrete SELFIES Molecular Generation via WGAN-GP with Gumbel-Softmax Relaxation

Faten F. Abushmmala[1]*, Mohammed A. Alhanjouri [2]

Type: Full Article. Received: 1st June 2026, Accepted: 16th July 2026

Accepted Manuscript, In Press

Abstract: This paper presents a deep generative framework for molecular design that tackles a key limitation of Generative Adversarial Networks (GANs) when applied to discrete molecular data. Although GANs have shown strong potential in continuous domains, their use in molecular sequence generation remains limited due to training instability and rapid mode collapse, even in advanced variants such as Wasserstein GAN with Gradient Penalty (WGAN-GP). In many cases, these models fail to produce meaningful or valid outputs. To address this issue, we combine Gumbel-Softmax relaxation with the SELFIES encoding scheme, allowing gradients to propagate effectively over discrete sequences. This design improves training stability and enables the model to avoid early collapse, making adversarial training practical in this setting. The model is trained on SMILES-based data converted to SELFIES encoding and evaluated using the GuacaMol distribution-learning benchmark. The results show high validity (97.25%), strong uniqueness (92.96%), and very high novelty (99.21%). Unlike standard GAN configurations that often fail to converge, the proposed approach maintains stable training and produces chemically meaningful and diverse molecules with competitive performance. Overall, the findings suggest that with appropriate design choices, GAN-based models can become a viable, energy-efficient alternative for discrete molecular generation, providing a practical step toward sustainable, fast-tracked generative models for drug discovery.



Keywords: Energy-efficient computing, Molecular generation, WGAN-GP, Gumbel-Softmax, SELFIES, Generative adversarial networks, Drug discovery, Green computing.

Introduction

De novo drug design, the targeted purpose of molecular discovery, is fundamentally challenging 1]. Traditional high-throughput screening and combinatorial chemistry consume immense physical and environmental resources, yet they frequently yield few viable drug candidates [23] Recently, deep generative models have shown promising potential to explore the massive chemical compound space efficiently, saving substantial laboratory resources and reducing chemical waste.

While our method builds on established components (WGAN-GP, Gumbel-Softmax, and SELFIES), this work makes three original contributions. First, we modify the critic so it can process both tokenized real sequences and the model's soft one-hot outputs using a single embedding pathway. Second, we introduce a gentle temperature annealing schedule ($1.2 \rightarrow 0.9$) that enables smooth discretization without destabilizing training. Third, we add an early-stopping rule based on molecular validity to avoid mode collapse and overtraining during adversarial generation of discrete sequences.

Several prior works have explored molecular generation using SELFIES and other deep generative frameworks, but they differ substantially in both purpose and evaluation. SELFormer [4] introduced a RoBERTa-based chemical language model trained on SELFIES representations from ChEMBL and later fine-tuned for molecular property prediction on MoleculeNet tasks. It reported strong results on regression and classification benchmarks, including ESOL and SIDER. However, SELFormer is primarily a property prediction model rather than a generative one, and it was not evaluated on distribution-learning benchmarks such as GuacaMol. For that reason, it is methodologically different from our work and cannot be treated as a direct baseline.

DiGress [5] is more relevant to our setting because it was evaluated on GuacaMol and targets molecular generation directly. It uses a discrete denoising diffusion process over molecular graphs, where node and edge categories are generated through iterative refinement with a graph transformer. The model achieved strong results on Guacamol, including high

[1] Department of Computer Engineering, Faculty of Engineering, Palestine Technical College of Deir Al-Balah (PTCD, Gaza, Palestine
E-mail: fshmmala@ptcdb.edu.ps
*Corresponding Author ORCID: https://orcid.org/0000-0002-8422-6116
[2] Department of Computer Engineering, Faculty of Engineering, Islamic University, Gaza, Palestine
E-mail: mhanjouri@iugaza.edu.ps , ORCID: https://orcid.org/0000-0002-7238-6614

validity, uniqueness, novelty, and competitive KL divergence and FCD scores. Even so, it differs from our approach in a key way: it works on graph-based molecular representations, whereas our model operates on SELFIES sequences. In addition, DiGress depends on a diffusion-based graph generation process with explicit modelling of atoms and bonds across multiple denoising steps, making it architecturally more complex and computationally heavier than our adversarial sequence-based framework.

The Regression Transformer [6] also sits somewhat outside our setting. Although it combines property prediction and property-conditioned molecular generation within a single framework, it was developed for a different goal and evaluated under a different protocol. Its results are reported on regression benchmarks and constrained optimization tasks, not on GuacaMol-style distribution-learning benchmarks. As a result, the reported metrics do not reflect the same aspects of generative quality that matter in our work, such as validity, uniqueness, novelty, KL divergence, and Fréchet ChemNet Distance. We therefore mention the Regression Transformer as a related generative direction, but not as a directly comparable baseline.

Overall, these studies show the breadth of recent work in molecular modelling, but they also highlight that our contribution occupies a different space: we focus specifically on adversarial generation of chemically valid, diverse, and distributionally realistic molecules using SELFIES under the GuacaMol evaluation framework.

Generative Adversarial Networks (GANs) [7] are powerful generative models that have been applied in drug design in multiple studies, particularly for generating novel molecules and compound structures. In the context of de novo molecular design and discovery, GANs generate entirely new molecules with certain desired properties, offering significant potential for drug discovery. A GAN consists of two competing neural networks: a generator and a discriminator. The generator aims to produce synthetic molecular structures that resemble real data, while the discriminator evaluates whether a given sample is real (from the dataset) or generated. Through this adversarial training process, the generator progressively improves its ability to produce realistic and chemically meaningful molecules.

Formally, the training of GANs is defined as a minimax optimization problem between the generator *G* and the discriminator *D*:

$$\min_G \max_D \mathbb{E}_{x\sim\mathcal{P}data(\mathcal{X})}[logD_{(\mathcal{X})}] + \mathbb{E}_{Z\sim\mathcal{P}_z(z)}\left[\log\left(1 - D\left(G(\mathcal{Z})\right)\right)\right] \quad (1)$$

where χ denotes real molecular samples drawn from the data distribution ρ_{date}, and z is a latent noise vector sampled from a prior distribution p_z. The generator *G(z)* maps the noise vector to a generated molecular structure, while the discriminator *D(χ)* outputs the probability that a sample is real. The generator aims to minimize this objective by producing samples that fool the discriminator, whereas the discriminator seeks to maximize it by correctly distinguishing real from generated molecules.

The Wasserstein GAN (WGAN) [8] introduces a new loss function based on the Earth Mover's (Wasserstein) distance. This distance metric provides a more stable gradient and is less prone to issues such as mode collapse. The Wasserstein distance measures the cost of transforming the distribution of generated samples into the distribution of real samples. The WGAN formulation replaces the discriminator with a critic network and enforces a Lipschitz continuity constraint, commonly implemented through weight clipping.

$$L = \mathbb{E}_{\mathcal{X}\sim\mathcal{P}_{real}}[\mathcal{F}(\mathcal{X})] - \mathbb{E}_{z\sim\mathcal{P}_z}[f(g(z))] \quad (2)$$

where *f* denotes the critic function, *g* represents the generator, and ρ_{real} and ρ_z correspond to the real data distribution and the noise distribution, respectively.

WGAN-GP 9] is an improvement over the original WGAN that replaces weight clipping with a gradient penalty to enforce the Lipschitz constraint. The gradient penalty term is added to the loss function to penalize deviations from 1 in the gradient norm of the critic function with respect to its input:

$$L = \mathbb{E}_{\mathcal{X}\sim\mathcal{P}_{real}}[f(\mathcal{X})] - \mathbb{E}_{z\sim\mathcal{P}_z}[f(g(z))] + \lambda\mathbb{E}_{z\sim\mathcal{P}_{\hat{x}}}[(||\nabla_{\hat{x}}f(\hat{x})||_2 - 1)^2 (3)$$

where $\hat{x}$ is an interpolation between real and generated samples.

In general, GANs have been widely used for data generation tasks. However, applying GANs to discrete molecular representations such as SMILES is challenging due to non-differentiability and strict syntax constraints. To address these limitations, this work proposes a hybrid framework that combines WGAN-GP with Gumbel-Softmax and SELFIES representation.

Furthermore, the proposed model is evaluated against the established GuacaMol benchmark [10], and the results are critically analyzed and discussed in comparison with the reported baseline performances. Table 1 provides a summary of the results reported in the GuacaMol benchmark [10], highlighting the performance of state-of-the-art models in de novo molecular design.

Table (1): Results of baseline models for the distribution-learning benchmarks.

Model	Validity	Uniqueness	Novelty	KL(exp(−DKL))	FCD
Random Sampler	1.000	0.997	0.000	0.998	0.929
SMILES LSTM (1)	0.959	1.000	0.912	0.991	0.913
Graph MCTS	1.000	1.000	0.994	0.522	0.015
AAE	0.822	1.000	0.998	0.886	0.529
ORGAN (1)	0.379	0.841	0.687	0.267	0.000
VAE	0.870	0.999	0.974	0.982	0.863

Materials and Methods

In this section, we first describe the dataset used in this study, which is obtained from the ChEMBL database [10]. It is important to note that this dataset is encoded using the SMILES representation. The choice of this dataset is motivated by its use in the GuacaMol benchmark, making it suitable for comparison with prior work. Next, we describe the process of converting the molecular representations from SMILES to SELFIES. Following that, we present the architecture of the proposed model, including the key components and hyperparameters. Finally, we discuss the evaluation metrics used to assess the performance of the model

Data preparation

The most widely used molecular representations in academic research are SMILES and SELFIES [11,12]. SMILES (Simplified Molecular Input Line Entry System) is a line notation for describing the structure of chemical species using short ASCII strings. For example, the SMILES for ethanol (C2H5OH) is CCO. SELFIES (Self-Referencing Embedded Strings) is a newer, robust molecular string representation designed to avoid invalid molecule representations, which can occur with SMILES.

Molecules are represented using SELFIES, which ensures syntactic validity. Sequences are tokenized and converted into integer indices. Special tokens such as <bos>, <eos>, and <pad> are included.

SMILES to SELFIES Conversion and Data Preprocessing

In this work, the molecular dataset is initially represented using the SMILES (Simplified Molecular Input Line Entry System) notation. In order to improve the robustness of sequence generation and reduce the likelihood of producing invalid molecular structures, the SMILES representations are converted into SELFIES (Self-Referencing Embedded Strings), which provide a 100% valid molecular string representation by construction. The conversion process is performed using the SELFIES encoder function, which maps each SMILES string into its corresponding SELFIES representation. Invalid or non-decodable SMILES strings that fail during the conversion process are discarded to ensure data consistency and quality.

About 2.3% of SMILES strings in the full ChEMBL dataset (around 3,500 entries) failed SELFIES conversion and were removed. Although this filtering is necessary for data quality, it may slightly shift the dataset by disproportionately excluding molecules with unusual substructures or rare functional groups that SELFIES struggles to encode. Future work should examine the structural features of the discarded molecules to assess any bias introduced by this step.

Following the conversion, each SELFIES string is tokenized into a sequence of discrete symbols. Unlike SMILES, where tokenization can be ambiguous, SELFIES provides a well-defined token structure in which each token corresponds to a chemically valid unit enclosed within brackets. This property simplifies the tokenization process and ensures consistency across the dataset. After extracting the tokens, a vocabulary is built by collecting all the unique SELFIES symbols found in the dataset. In addition to these, three special tokens are added: <pad> for padding sequences, <bos> to mark the beginning of a sequence, and <eos> to indicate its end. Each token is then assigned a unique numerical index, allowing the sequences to be represented in a format suitable for model training. Subsequently, each tokenized SELFIES sequence is encoded into a sequence of integer IDs based on the constructed vocabulary. The special tokens <bos> and <eos> are added to the beginning and end of each sequence, respectively. To enable batch processing, all sequences are padded or truncated to a fixed maximum length, determined as the 99th percentile of sequence lengths in the dataset. This approach balances computational efficiency while preserving most of the structural information in the data. The final output of this preprocessing pipeline is a matrix of token IDs representing the SELFIES sequences, along with the corresponding vocabulary mapping. These processed representations are used as input to train the generative model.

Model Architecture

The proposed molecular generation framework is based on a Wasserstein Generative Adversarial Network with Gradient Penalty (WGANGP), adapted for discrete sequence generation using the Gumbel- Softmax relaxation. The model is designed to learn the distribution of molecular strings represented in the SELFIES format and to generate novel molecular sequences in a stable adversarial training setting. The framework is built around two main parts: a generator and a critic. The generator takes as input a random noise vector z ∈ R128, which is sampled from a standard normal distribution, and maps it into a sequence of token logits corresponding to the molecular vocabulary. Architecturally, the generator starts with a fully connected layer of size 256 using ReLU activation, which maps the input noise vector into a higher-level representation. This representation is then repeated across all time steps using a repeat vector, creating a fixed-length sequence that can be further processed. The repeated representation is processed by a gated recurrent unit (GRU) layer with 256 hidden units and return sequences enabled, allowing the generator to model sequential dependencies between molecular tokens. Finally, a dense output layer projects the hidden states at each time step into the vocabulary space, producing a sequence of logits over the SELFIES token set.

Since molecular sequences are discrete, direct backpropagation through sampled tokens is not possible. To address this issue, the Gumbel-Softmax trick with the straight-through estimator is employed. During training, the generator output logits are converted into differentiable approximations of one-hot vectors, enabling gradient flow through the discrete sampling process while preserving the token-level structure required for adversarial learning.

The critic is designed to distinguish between real molecular sequences from the training dataset and generated sequences produced by the generator. It receives either tokenized real sequences or soft/hard onehot generated sequences. For discrete token inputs, the critic first applies an embedding layer of dimension 64 to map tokens into dense vector representations. For generated one-hot-like inputs, the critic directly computes the embedded representation through matrix multiplication with the embedding weights. The embedded sequence is then passed through a dropout layer with a dropout rate of 0.3 to improve regularization, followed by a GRU layer with 128 hidden units to capture sequential structural information. A second dropout layer is applied before a final dense layer outputs a scalar critic score. The training procedure follows the WGAN-GP formulation. Instead of the standard GAN discriminator objective, the critic is optimized using the Wasserstein loss, which provides more stable gradients for sequence generation tasks. To enforce the Lipschitz continuity constraint required by Wasserstein training, a gradient penalty term is added to the critic loss. This penalty is computed on interpolated samples between real and generated data, and the gradient norm is encouraged to remain close to 1. In this work, the gradient penalty coefficient is set to λGP = 5.0.

To further stabilize adversarial learning, the critic is updated more frequently than the generator. Specifically, for each generator update, the critic is trained for $n_{critic} = 3$ iterations. The generator is optimized to maximize the critic's score on generated samples, whereas the critic is trained to assign higher scores to real sequences and lower scores to generated ones. The Adam optimizer is used for both networks, with separate learning rates to balance the adversarial game: 2 × 10−4 for the generator and 1 × 10−4 for the critic, with $\beta 1$ = 0.0 and $\beta 2$ = 0.9. The training dataset consisted of exactly 147,634 tokenized SELFIES sequences that after preprocessing and filtering. The model is trained using mini-batches of size 64 for 40 epochs, with 200 training steps per epoch. In order to improve sampling behavior during training, the Gumbel-Softmax temperature is annealed gradually from 1.2 to 0.9 across epochs. This relatively mild annealing schedule allows the model to transition smoothly from soft probabilistic token selections to sharper discrete approximations without causing abrupt instability. To prevent mode collapse and overtraining, an early stopping strategy is incorporated based on molecular validity. At the end of each epoch, the generator is evaluated by sampling 300 molecular sequences, which are decoded from SELFIES into SMILES and tested for syntactic validity. The best-performing generator checkpoint is saved whenever the validation validity score improves. Training is terminated if no improvement is observed for five consecutive epochs. Overall, the proposed architecture combines sequential neural modeling,

differentiable discrete sampling, and Wasserstein-based adversarial optimization to provide a robust framework for molecular generation in the SELFIES representation space.

Evaluation Metrics

The performance of the proposed model was assessed using both validity- and distribution-based evaluation metrics [13], [3], following the GuacaMol distribution-learning protocol. These metrics were selected to measure not only whether the generated molecular strings correspond to chemically valid structures, but also whether the generated set is diverse, novel, and statistically aligned with the training distribution.

Validity measures the proportion of generated molecules that can be successfully decoded and parsed into chemically valid molecular structures. In this work, generated SELFIES strings were first decoded into SMILES and then validated using RDKit. This metric reflects the syntactic and structural correctness of the generated molecules.

Uniqueness quantifies the proportion of non-duplicate molecules among the valid generated samples. It is computed after canonicalization, ensuring that equivalent molecular structures represented in different textual forms are counted only once. This metric measures how varied the generated molecules are, instead of producing the same ones repeatedly.

Novelty measures how many of the generated molecules are new and not seen in the training data. It helps check whether the model is creating new structures instead of just repeating what it has already learned.

KL divergence was used to compare the distribution of selected physicochemical properties between the generated molecules and the training set. Specifically, the comparison was performed over common molecular descriptors, including molecular weight (MW), LogP, topological polar surface area (TPSA), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), number of rotatable bonds (RotB), and ring count. For each property, the Kullback‑ Leibler divergence was computed between the histogram-based distributions of the generated and reference sets. Following the GuacaMol benchmark, an aggregated score was then obtained as the mean of exp(−DKL) across all evaluated properties, so that higher values indicate better agreement with the reference distribution.

Internal diversity was used to evaluate structural diversity within the generated molecular set. This metric was estimated from pairwise Tanimoto similarities computed on Morgan fingerprints, and reported as one minus the average similarity. Higher internal diversity values indicate a more structurally varied set of generated molecules.

Frechet ChemNet Distance (FCD) was used as an additional distributional similarity metric between generated molecules and the reference dataset. FCD is conceptually similar to the Frechet Inception Distance used in image generation, but adapted to molecular data through ChemNet embeddings. Lower raw FCD values indicate closer agreement between generated and reference molecule distributions. We report the raw FCD value directly to allow straightforward comparison with the literature. A GuacaMol-style normalized score was also computed, but only for reference. Together, these metrics provide a comprehensive evaluation of molecular generation quality, covering chemical validity, sample diversity, novelty, and similarity to the target data distribution.

Experimental Setup

The experiments were conducted using the proposed SELFIES-based WGAN-GP model with Gumbel-Softmax for differentiable discrete sequence generation on Colab Pro+ (Python). The training dataset contained exactly 147,634 tokenized SELFIES sequences, obtained from the molecular dataset after preprocessing. During training, the data were shuffled using a buffer size of 30,000 and processed in minibatches of size 64. The generator received a latent noise vector of dimension 128 sampled from a standard normal distribution.

The model was trained for up to 40 epochs, with 200 training steps per epoch. To stabilize adversarial learning, the critic was updated three times for each generator update, following the standard WGAN training strategy. The gradient penalty coefficient was set to 5.0 in order to enforce the Lipschitz constraint while avoiding excessively strong critic regularization. The optimization process employed the Adam optimizer for both networks, with a learning rate of 2×10^{-4} for the generator and 1×10^{-4} for the critic, and momentum parameters $\beta_1 = 0.0$ and $\beta_2 = 0.9$. These settings were chosen to keep the training stable and to maintain a reasonable balance between the generator and the critic. To further reduce instability, gradient clipping with a maximum norm of 5.0 was applied during both generator and critic updates. Since the model works with discrete molecular sequences, Gumbel- Softmax was used during training. The temperature was gradually reduced from 1.2 to 0.9 over the epochs, allowing the model to move smoothly from softer probability distributions to more discrete outputs without sudden changes in training behavior. The key hyperparameters were selected as follows. The learning rates and β values follow established WGAN-GP conventions from Gulrajani et al. [9].

Results and Discussion

In this section, we present the evaluation of the proposed SELFIES based WGAN-GP model using the GuacaMol benchmark. The model's performance is analyzed across five standard distribution-learning metrics: validity, uniqueness, novelty, Kullback–Leibler (KL) divergence, and Frechet ChemNet Distance (FCD).

The model achieved a chemical validity score of 97.25%, demonstrating that the combination of the SELFIES representation and the Gumbel-Softmax relaxation successfully addresses the validity issues typically associated with GAN-based molecular generation. By mapping the continuous outputs of the generator to differentiable approximations of one-hot vectors, the framework allows the model to learn valid syntactic structures without collapsing into invalid string configurations. The small fraction of invalid sequences (2.75%) is primarily due to truncated strings where the <eos> token was generated prematurely or where the padding length cut off a long molecular sequence.

In terms of sample diversity, the model reached a uniqueness score of 92.96% and an internal diversity score of 0.884. This indicates that the generator avoids mode collapse, a common failure mode in generative adversarial training, and is capable of exploring a wide region of the chemical space. The competitive novelty score of 99.21% further confirms that the model does not merely memorize the training dataset, but instead generalizes from the learned distribution to propose novel molecular structures.

To evaluate how well the generated molecules mirror the underlying training data, we calculated the KL divergence across

several key physicochemical properties and the raw FCD. The results are summarized in Table 2 and compared against the baseline models reported in the GuacaMol benchmark.

Table (2): Experimental results of the proposed framework compared to GuacaMol baselines

Model	Validity	Uniqueness	Novelty	KL Score	FCD
SMILES LSTM (1)	0.959	1.000	0.912	0.991	0.913
VAE	0.870	0.999	0.974	0.982	0.863
ORGAN (1)	0.379	0.841	0.687	0.267	0.000
Proposed WGAN-GP (SELFIES)	**0.973**	**0.930**	**0.992**	**0.814**	**0.442**

The proposed model significantly outperforms ORGAN, which is a standard RL-based GAN applied to SMILES, in all categories. Specifically, ORGAN suffers from low validity (37.9%) due to the discrete nature of SMILES tokens and the discrete rewards used during training. In contrast, our framework achieves 97.25% validity, illustrating the clear advantage of using the robust SELFIES encoding combined with a continuous relaxation technique. While the SMILES LSTM and VAE models show higher structural alignment with the training distribution (reflected in higher KL and FCD scores), our model shows a higher novelty rate (99.21%) compared to the LSTM (91.2%) and VAE (97.4%). This indicates that while the adversarial training setup might introduce a slight distributional shift, it encourages the generator to explore unique regions of the chemical space, making it a valuable tool for de novo drug discovery where discovering entirely novel scaffolds is a primary goal.

Conclusion

In this work, we developed and evaluated a deep generative framework for molecular design that combines Wasserstein Generative Adversarial Networks with Gradient Penalty (WGAN-GP), the Gumbel-Softmax relaxation, and the SELFIES molecular string representation. Our main goal was to address the training instabilities and low validity rates that traditionally limit the use of GANs in discrete sequence generation.

The experimental results on the GuacaMol distribution-learning benchmark demonstrate that our approach provides a stable training regime and generates high-quality molecules. The model achieved a validity score of 97.25%, a uniqueness score of 92.96%, and a novelty score of 99.21%, significantly outperforming older GAN-based molecular models like ORGAN. These findings demonstrate that when paired with a robust molecular representation like SELFIES and an appropriate continuous relaxation method, GAN-based architectures can become reliable and effective tools for exploring the chemical compound space, which aligns with recent sustainable computing paradigms and green chemical evaluations published in the literature [14]. Future work will focus on integrating property-conditional generation into this framework, allowing the model to optimize specific chemical or pharmacological properties, such as binding affinity or synthetic accessibility, alongside distribution learning.

Disclosure Statement

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Research source

This study is derived from my PhD dissertation in Islamic University-Gaza. This manuscript forms part of the PhD dissertation of Faten Abushmmala at the Islamic University of Gaza. The dissertation is currently in the defense and publication process and has not yet been deposited in a publicly accessible repository

Availability of data and materials

The raw data required to reproduce these findings are available in the body and illustrations of this manuscript.

Author's contribution

The authors confirm contribution to the paper as follows: study conception and design: F. F. Abushmmala, M. A. Alhanjouri; software programming, data generation, and drafting of the manuscript: F. F. Abushmmala; formal analysis, critical review, and supervision: M. A. Alhanjouri. All authors reviewed the results and approved the final version of the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this article

Acknowledgements

The authors would like to thank Islamic University for their support in completing this work.

References

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. 2016 May 1;47:20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

2. Schneider G. Automating drug discovery. Nat Rev Drug Discov. 2018 Feb;17(2):97–113. https://doi.org/10.1038/nrd.2017.232

3. Schneider G, Clark DE. Automated De Novo Drug Design: Are We Nearly There Yet? Angew Chem Int Ed. 2019 Aug 5;58(32):10792–803. https://doi.org/10.1002/anie.201814681

4. Yüksel A, Ulusoy E, Ünlü A, Doğan T. SELFormer: molecular representation learning via SELFIES language models. Mach Learn Sci Technol. 2023 Jun 1;4(2):025035. https://doi.org/10.1088/2632-2153/acdb30

5. Vignac C, Krawczuk I, Siraudin A, Wang B, Cevher V, Frossard P. DiGress: discrete denoising diffusion for graph generation. In: International Conference on Learning Representations (ICLR) [Internet]. 2023. https://openreview.net/forum?id=UaAD-Nu86WX

6. Born J, Manica M. Regression Transformer enables concurrent sequence regression and generation for molecular language modelling. Nat Mach Intell. 2023 Apr 6;5(4):432–44. https://doi.org/10.1038/s42256-023-00639-z

7. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS) [Internet]. 2014. p. 2672–80. https://arxiv.org/abs/1406.2661

8. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML) [Internet]. 2017. p. 214–23. https://proceedings.mlr.press/v70/arjovsky17a.html

9. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems (NeurIPS) [Internet]. 2017. p. 5767–77. https://arxiv.org/abs/1704.00028

10. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking Models for de Novo Molecular Design. J Chem Inf Model. 2019 Mar 25;59(3):1096–108. https://doi.org/10.1021/acs.jcim.8b00839

11. 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 Dec 1;1(4):045024. https://doi.org/10.1088/2632-2153/aba947

12. Wiswesser WJ. 107 years of line-formula notations (1861-1968). J Chem Doc. 1968;8(3):146–50. https://doi.org/10.1021/c160030a004

13. 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 May 23;25(4):bbae338. https://doi.org/10.1093/bib/bbae338

14. Hawash M, Jaradat N, Abualhasan M, Qaoud MT, Joudeh Y, Jaber Z, et al. Molecular docking studies and biological evaluation of isoxazole-carboxamide derivatives as COX inhibitors and antimicrobial agents. 3 Biotech. 2022 Dec;12(12):342. https://doi.org/10.1007/s13205-022-03408-8