Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 15 No. 4, 2024 - www.cisdijournal.net

LAIgnd: Revolutionizing Drug Discovery with Advanced AI-Driven Molecule Generation

[1]*Obi, E.D., [1]Yentumi, J.A., [1]Mbatuegwu, D., [2]Omotuyi, O.I., [3]Ajayi, O.O. & [1]Nwokoro, A.
[1]Autogon Inc. 3002 Falls at Fairdale, 77057, USA
[2]Department of Pharmacology and Toxicology, Afe Babalola University, Ado-Ekiti, Nigeria
[3] Department of Computer Science, Adekunle Ajasin University, Nigeria
E-mail: eobi@autogon.ai, joshua@autogon.ai, david@autogon.ai, olaposi.omotuyi@abuad.edu.ng, olusola.ajayi@aaua.edu.ng, aaron@autogon.ai;
Phone: [1*]+1 832 925 1036, [1]+233 54 100 6410, [1]+234 816 873 2219, [2]+234 807 094 3256, [3]+27 64 0499906, [1]+234 813 793 6228;

ABSTRACT

De novo molecular generation is crucial for advancing drug discovery and chemical research. This accelerates the search for new drug candidates and deepens our understanding of molecular diversity. The development of deep learning has propelled and expedited the de novo molecular generation. Generative networks, particularly Variational Autoencoders (VAEs), can randomly produce new molecules and modify molecular structures to enhance specific chemical properties, which are essential for advancing drug discovery. Although VAEs offer numerous advantages, they are hindered by limitations that affect their capacity to optimize properties and decode syntactically valid molecules. To address these challenges, we present LAIgnd, a de novo drug molecule generation model that implements a custom β-CVAE architecture conditioned on protein sequences and SELFIES input. Extensive experiments have shown that LAIgnd generates a wide variety of valid, novel, and effective molecules for complex and simple diseases, demonstrating its robustness and generalization capabilities. Additionally, by employing molecular docking, toxicity, similarity, and synthetic accessibility experiments, we demonstrated the drug-likeness and effectiveness of the generated molecules. The ability of our model to generate novel and diverse compounds was illustrated by a case study focusing on Lung Cancer. A total of four hundred (400) molecules were generated by LAIgnd, with a high number of molecules exhibiting strong inhibitory activity against the Epidermal Growth Factor receptor, as indicated by binding affinities. LAIgnd provides new insights into future directions to enhance therapeutics for complex and simple diseases by generating high-quality multi-target molecules for drug discovery.

Key words: De novo molecular generation, Drug discovery, Variational Autoencoders (VAEs), SELFIES, Protein sequences.

CISDI Journal Reference Format
Obi, E.D., Yentumi, J.A., Mbatuegwu, D., Omotuyi, O.I., Ajayi, O.O. & Nwokoro, A. (2024): LAIgnd: Revolutionizing Drug Discovery with Advanced AI-Driven Molecule Generation. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 15 No 4, Pp 1-10.
Available online at www.isteams.net/cisdijournal. dx.doi.org/10.22624/AIMS/CISDI/V15N3P4

1. INTRODUCTION

The field of molecular design and generation has undergone significant advancements in recent years, driven by the pressing need to efficiently explore vast chemical spaces and discover novel compounds with desired properties. (Meyers et al., 2021; Scannell & Bosley, 2016).

This pursuit is particularly crucial in drug discovery, in which innovative solutions are constantly sought to address complex challenges (Iwata et al., 2023). At the heart of this endeavor lies the concept of de novo molecular design, a sophisticated approach that leverages artificial intelligence to propose new chemical structures tailored to specific molecular profiles (Ang et al., 2023; Meyers et al., 2021). This method, also known as "generative chemistry," has gained traction owing to the increasing prevalence of AI-powered generative models in the field (Ai et al., 2024; Richards & Groener, 2022; Xue et al., 2019).

The sheer scale of the chemical space, encompassing all possible molecules, presents both a challenge and an opportunity (Ai et al., 2024; Cheng et al., 2021). Although traditional compound libraries, even those containing billions of molecules, represent only a fraction of this space, de novo design methods offer a more targeted and efficient approach to traversing this vast landscape (Richards & Groener, 2022; Meyers et al., 2021; Xue et al., 2019). By generating compounds in a directed manner, researchers aim to identify optimal chemical solutions while evaluating fewer molecules than traditional brute-force screening methods (Meyers et al., 2021). Recent developments in deep learning and reinforcement learning have further revolutionized the field of molecular generation.

These approaches enable the creation of novel molecules with tailored properties, accelerating drug development processes and expanding our understanding of molecular diversity (Ai et al., 2024; Ang et al., 2023; Li et al., 2021; Xue et al., 2019). The integration of various molecular representation techniques, such as SMILES, SELFIES (Meyers et al., 2021; Krenn et al., 2020; Gupta et al., 2018), and graph-based models, has enhanced the precision and robustness of these generative systems (Ang et al., 2023; Jin et al., 2018). Deep generative models such as conditional variational autoencoders (CVAEs) have demonstrated significant efficacy in the generation of molecules with specific properties. Nonetheless, these model architectures often encounter challenges in effectively segregating latent molecular representations, potentially resulting in posterior collapse, as observed in a standard VAE (Lee & Min, 2022; Richards & Groener, 2022; Higgins et al., 2017).

We introduce LAIgnd, which implements a deep learning architecture that relies on a custom conditional variational autoencoder (β-CVAE) model to generate SELFIES conditioned on protein sequences. The β-CVAE model combines powerful encoding and decoding components to generate SELFIES representations conditioned on protein sequences. The unique design of the model makes it a valuable tool for researchers and practitioners in the fields of molecular generation and drug discovery. The model pipeline also implements rigorous molecular validity tests and checks to ensure that the generated molecules meet the clinical requirements.

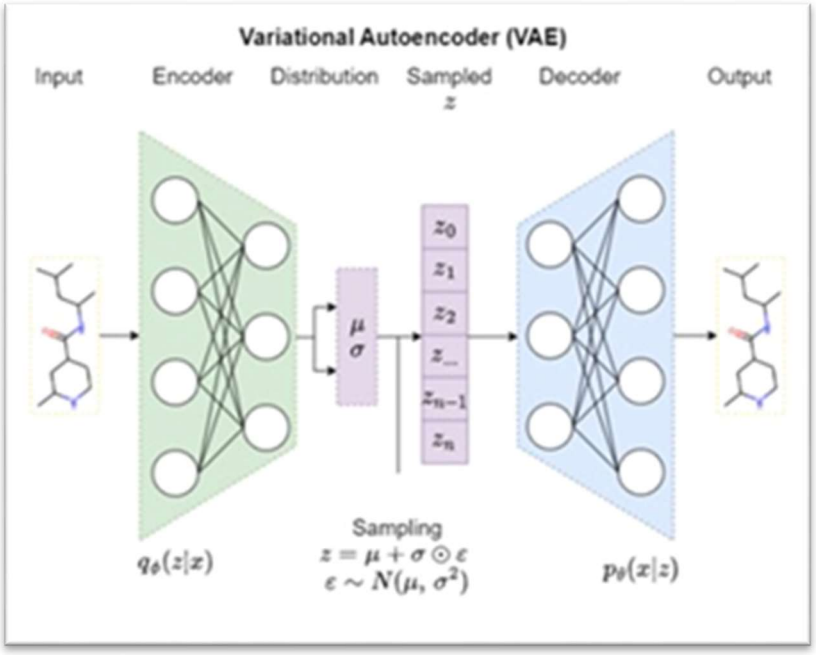

Fig. 1(a)

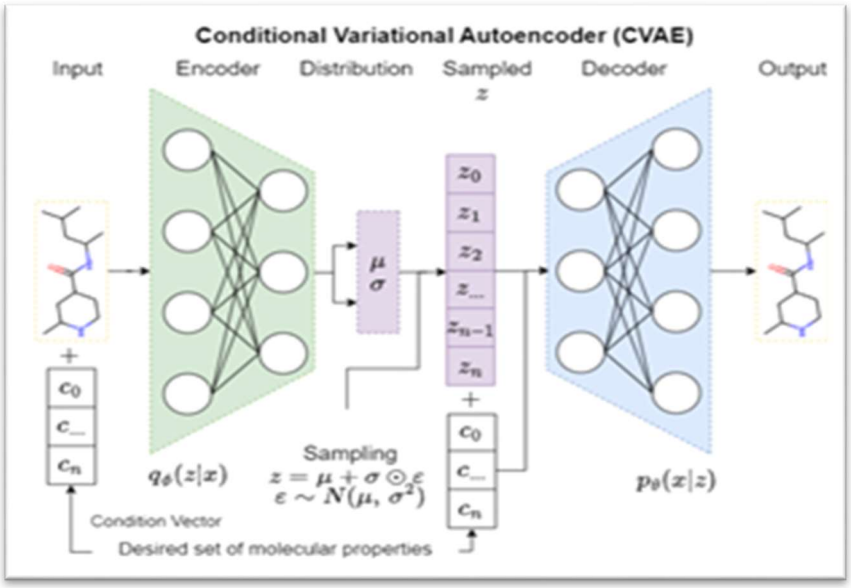

Fig. 1(b)

Fig. 1.(A & B) Differentiating between a standard VAE and a CVAE. The standard VAE (left) generates a new molecule similar to the input. The CVAE (right) generates a new molecule with the desired properties conditioned by the conditional vector input (Ang et al., 2023).

2. METHODOLOGY

2.1 Dataset
The dataset used to train the model was curated in-house.

2.2 β-Conditional Variational Autoencoder (β-CVAE) Architecture

The Conditional Variational Autoencoder (CVAE) model architecture utilized for sequence generation consists of two primary components: the encoder and the decoder network cells. The encoder processes tokenized input sequences for both SELFIES and protein sequences, transforming them into a continuous multi-dimensional vector through standard embedding layers. Long short-term memory cells are employed to capture contextual information, enabling the modeling of complex dependencies within the protein sequences. The output from these encoders is then fed into a combined encoder to generate the final latent representation. This representation is subsequently used to predict the mean and log variance of the latent distribution.

The decoder generates a SELFIES output based on the sampled latent vector and the protein sequence. The final layer employs a dense layer with softmax activation to predict the probability distribution over the SELFIES vocabulary. The model is trained using a combination of Kullback-Leibler (KL) divergence loss and cross-entropy loss to compare the predicted and target SELFIES representations. We introduced a hyperparameter β to mitigate the posterior collapse problem during training.

2.3. Inputs for Model Prediction
To test the real-world applicability of our model for drug discovery and molecule generation, we selected a global disease of concern. For this target, the protein sequence was retrieved from the National Center for Biotechnology Information (NCBI), cross-referencing with the UniProt Protein Database. The sequence was then fed to the model for molecule generation.

2.4. Post-Training Model Validations and Testing
The generated SMILE sequences were validated using the RDKit Chemoinformatics library. For each generated molecule, RDKit determined the validity of the molecule, calculated its molecular weight, and generated its molecular formula. Additionally, in-house docking and the ADMETox pipeline were utilized to validate the accuracy and validity of the molecules generated by the model.

2.4.1 Docking and Toxicity Test
The crystal structure of the retrieved target protein sequence was obtained from the Protein Data Bank (PDB), and molecular docking was performed using a double docking mechanism involving an initial docking step to determine the best poses, accomplished using the DiffDock molecular docking program. The docking results obtained from the initial docking were then passed to the second docking suite (GNINA) to calculate the binding affinities of the poses. A GNINA minimized affinity score threshold of ≤ -5 and a DiffDock confidence threshold of > -1 were used as filtering criteria to select the best molecules with potential activity against the target protein. The resulting filtered molecules were then fed into the in-house AdmetAI ADMETox pipeline, and a synthetic toxicity check was performed for the generated molecules.

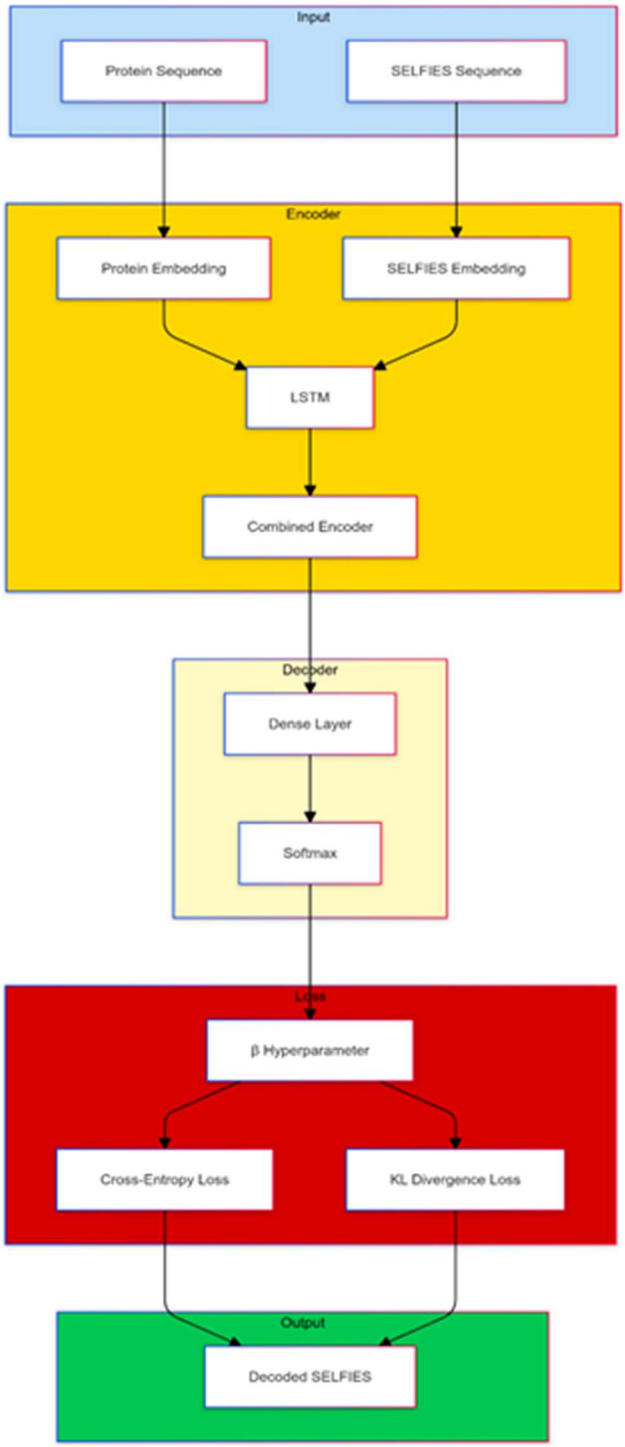

Fig. 2: LAIgnd Architecture Diagram

2.4.2 Tanimoto Similarity Test

We further implemented a similarity score check using the Tanimoto similarity function in the RDKit Chemoinformatics library to compare the similarity of the generated molecules with existing molecules. The similarity pipeline involves two major chemical molecule databases, PubChem and ChEMBL, with the molecule similarity check limited to the top three similar molecules from these databases. A custom similarity score was also calculated for each molecule, and all molecules were ranked based on the custom, Tanimoto, and synthetic accessibility scores. These metrics, coupled with the synthetic accessibility score or value (SA_score), provided an overview of the synthetic pathway for molecule generation and ease of synthesis.

2.4.3 Synthetic Accessibility Score (SA score)

To judge the hardness or softness of the molecules generated for chemical synthesis, a synthetic accessibility score was calculated for each hit molecule selected from the docking pipeline. This metric score, provided by the RDKit Chemoinformatics library, determines the ease of synthesis of a molecule. The score ranged from one (1) for very easy to synthesize) to 10 (very difficult to synthesize).

2.5 Report Generation

An interactive HTML-enabled report detailing hit candidates and their respective results was generated as the final output of the discovery pipeline.

3. RESULTS

3.1 Application of LAIgnd Predictive Power on Real Disease Proteins

The main goal of developing LAIgnd is to generate de novo valid drug molecules against undruggable disease targets for personalized drug development. To test the efficacy of this model, we tested it against the epidermal Growth Factor Receptor implicated in Lung Cancer. In total, (400) molecules were generated for this target. After passing through our rigorous testing and validation pipelines, three hundred and sixty-seven molecules showed moderate to high binding affinities for the target protein (sample shown in Table 1).

The RDKit validity synthetic accessibility score (SA score) revealed that approximately sixty-five (65%) of the molecules had moderate difficulty while twenty-six (26%) percent and nine (9%) were hard and soft, respectively. This metric allows our system to provide a quick snapshot for scientists to immediately measure the ease of synthesis of the hit molecules. A comprehensive report that provides a detailed overview of the top small-molecule candidates was generated as the final output of the pipeline.

Table 1: DiffDock-Gnina Docking Results of top 10 LAIgnd-generated small molecules

hit_molecules	diffdock_confidence	gnina_minimized_affinity	molecular_weight	SA_Score	SA_Score_label
1	-0.90	-6.57926	142.206	4.6162	Moderate
2	-0.92	-5.61831	143.146	4.9414	Moderate
3	-0.79	-5.94425	304.283	5.9135	Moderate
4	-0.71	-5.60442	265.341	5.9585	Moderate
5	-0.20	-5.53803	270.293	6.3784	Difficult
6	-0.95	-5.97381	189.179	6.2976	Difficult
7	-0.62	-5.19615	145.122	6.2751	Difficult
8	-0.91	-5.46924	242.275	6.0094	Moderate
9	-0.87	-6.32216	295.428	6.8766	Difficult
10	-0.52	-4.91387	171.284	3.8971	Easy

For the Tanimoto similarity, two of the candidates received a high score of twenty-seven (27%). This was followed by two molecules with twenty-five (25%) percent similarity, and ten molecules with similarities ranging between twenty-four (24%) and nineteen (19%) percent respectively, showcasing the novelty of molecules generated by LAIgnd. Our pipeline also included a Pan-Assay Interference Compound (PAINS) score to screen molecules that may yield false-positive results during high-throughput screening assays (Table 2). These compounds often interfere with many biological assays, regardless of the protein target, and are crucial for avoiding false leads in the drug discovery process. Ninety-four percent (94%) of our molecules had a negative score (false) for PAINS, further validating the targeted nature of the molecules generated by our system.

Table 2: Synthetic ADMETox Screening results of the top 10 LAIgnd-generated small molecules

Hit Molecule	Epoxide Ring Present	PAINS	logP	Lipinski	QED	AMES	BBB Martins	Bioavailability Ma	Carcinogens Lagunin	ClinTox
1	False	False	-0.6464	4	0.3935	0.9871	0.5820	0.9600	0.4491	0.0800
2	False	False	-0.5894	4	0.4519	0.9232	0.7280	0.9379	0.6513	0.0750
3	False	False	1.027	4	0.3104	0.8755	0.8615	0.9714	0.5052	0.2991
4	False	False	0.19269	4	0.3083	0.9700	0.5527	0.9656	0.7095	0.3690
5	False	False	-0.17635	4	0.6976	0.9207	0.7202	0.9596	0.3959	0.2157
6	False	False	-1.6498	4	0.3664	0.9990	0.3516	0.9548	0.9448	0.0572
7	False	False	-2.8392	4	0.2940	0.9993	0.4901	0.9889	0.8193	0.1456
8	False	False	0.5693	4	0.6266	0.8235	0.9401	0.9656	0.2016	0.1530
9	False	False	2.90479	4	0.6400	0.9466	0.7137	0.8905	0.6194	0.4183
10	False	False	1.9305	4	0.7012	0.1797	0.9913	0.9461	0.1984	0.0632

4. DISCUSSION

β-CVAE architecture for molecule generation. Our system allows sampling from different latent spaces with each prediction run, thus improving the degree of novelty of the generated molecules. Furthermore, owing to the limitations of SMILES, we also introduced the SELFIES representation of the molecule to enhance the accuracy of the generated molecules. To evaluate the quality of the generated molecules, we implemented custom screening tests using industry-standard tools such as RDKit and structural similarity checks on large-scale chemical molecule databases, such as PubChem and ChEMBL. Our results give further credence to the novelty of molecules generated by our model by giving the highest similarity score of 27%. The evaluation of LAIgnd, a novel approach in molecular design for disease treatment, faces two primary obstacles. In the realm of machine learning, especially for tasks involving the generation of new molecular entities targeting multiple disease proteins, there is a notable absence of established benchmark datasets. This gap hinders the ability to objectively assess and compare the performance of LAIgnd against other models or established standards.

Additionally, the molecules proposed by LAIgnd are, by design, novel entities. This novelty, while potentially groundbreaking, introduces significant hurdles in the validation process including but not limited to, synthesis requirements and resource intensiveness. Before any biological testing can occur, these new molecules must be synthesized in a laboratory setting. The synthesis process demands specialized knowledge, considerable time, and substantial financial investment. These factors collectively impede the swift and comprehensive validation of LAIgnd's efficacy in producing therapeutically promising molecules. The situation underscores a broader challenge in the field of AI-driven drug discovery; balancing innovative potential with practical constraints in validation and implementation.

By conducting docking and ADMETox screening, we established drug-likeness and inhibitory validation of the generated molecules against the input target protein sequence. However, our model has room for improvement. Overall, LAIgnd presents continuous efforts to innovate the healthcare industry through artificial intelligence.

5. CONCLUSION

In summary, we implemented a β-CVAE deep generative model to generate novel drug molecules. We introduced an end-to-end pipeline that enables efficient generation and validation of potential small molecules for drug applications. Our results showed that our model is capable of generating unique molecules with optimized drug-related properties. By conditioning on the protein sequence input, LAIgnd provides valuable insights into personalized medicine using artificial intelligence. As we continue to improve the model, we will implement more stringent molecular validation methods and training parameters, further enhancing its potential application in the pharmaceutical and healthcare industries.

Image Redaction Statement

Owing to potential commercial interests, Autogon Inc. reserves the right to redact images of molecules generated by LAIgnd from this research report. This redaction is intended to protect proprietary information while maintaining the integrity of research findings.

Funding
This work was funded by Autogon Inc.

Credit Author Statement
Conceptualization: Obi E, Mbatuegwu D, Yentumi J. Data curation: Yentumi J, Mbatuegwu D. Formal analysis: Obi E, Yentumi J, Mbatuegwu D. Investigation: Obi E, Yentumi J, Mbatuegwu D, Omotuyi O. Methodology: Yentumi J, Mbatuegwu D, Obi E. Project administration: Obi E, Yentumi J, Mbatuegwu D. Resources: Obi E, Yentumi J, Mbatuegwu D, Nwokoro A. Software: Obi E, Yentumi J, Mbatuegwu D, Nwokoro A. Supervision: Obi E, Yentumi J, Mbatuegwu D. Validation: Obi E, Yentumi J, Mbatuegwu D, Nwokoro A. Visualization: Yentumi J, Mbatuegwu D. Writing – original draft: Yentumi J. Writing – review, and editing: Yentumi J, Obi E, Mbatuegwu D, Omotuyi O.

Declaration of competing interest
The authors declare that they are in partnership with Autogon Inc.

Code availability
The code employed in this application development is a proprietary asset that is not available to the public.

Acknowledgments
Support from Autogon management and the entire team is appreciated.

REFERENCES

1. Ai, C., Yang, H., Liu, X., Dong, R., Ding, Y., & Guo, F. (2024). MTMol-GPT: De novo multi-target molecular generation with transformer-based generative adversarial imitation learning. PLoS Computational Biology, 20(6 June). https://doi.org/10.1371/journal.pcbi.1012229
2. Ang, G. J. N., Chin, D. T. I., & Shen, B. (2023). Balancing Exploration and Exploitation: Disentangled β-CVAE in De Novo Drug Design. http://arxiv.org/abs/2306.01683
3. Cheng, Y., Gong, Y., Liu, Y., Song, B., & Zou Q. (2021). Molecular design in drug discovery: a comprehensive review of deep generative models. Briefings in Bioinformatics. 22(6):bbab344. https://doi.org/10.1093/ bib/bbab344 PMID: 34415297
4. Gupta, A., Müller, A. T., Huisman, B. J. H., Fuchs, J. A., Schneider, P., & Schneider, G. (2018). Generative recurrent networks for de novo drug design. Molecular informatics, 37(1-2):1700111.
5. Iwata, H., Nakai, T., Koyama, T., Matsumoto, S., Kojima, R., & Okuno, Y. (2023). VGAE-MCTS: A New Molecular Generative Model Combining the Variational Graph Auto-Encoder and Monte Carlo Tree Search. Journal of Chemical Information and Modeling, 63(23), 7392–7400. https://doi.org/10.1021/acs.jcim.3c01220
6. Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. In International conference on machine learning, pages 2323–2332. PMLR.
7. Krenn, M., Häse, F., Nigam, A. K., Friederich, P., & Aspuru-Guzik, A. (2022). Self-referencing embedded strings (selfies): A 100% robust molecular string representation. Machine Learning: Science and Technology, 1(4):045024, 2020

8. Lee, M., & Min, K. (2022). MGCVAE: Multi-Objective Inverse Design via Molecular Graph Conditional Variational Autoencoder. Journal of Chemical Information and Modeling, 62(12), 2943–2950. https://doi.org/10.1021/acs.jcim.2c00487
9. Li, Y., Pei, J., & Lai, L. (2021). Structure-based: De novo drug design using 3D deep generative models. Chemical Science, 12(41), 13664–13675. https://doi.org/10.1039/d1sc04444c
10. Meyers, J., Fabian, B., & Brown, N. (2021). De novo molecular design and generative models. In Drug Discovery Today (Vol. 26, Issue 11, pp. 2707–2715). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2021.05.019
11. Richards, R. J., & Groener, A. M. (2022). Conditional β-VAE for De Novo Molecular Generation. http://arxiv.org/abs/2205.01592
12. Scannell, J. W., & Bosley, J. (2016). When quality beats quantity: Decision theory, drug discovery, and the reproducibility crisis. PLoS ONE, 11(2). https://doi.org/10.1371/journal.pone.0147215
13. Xue, D., Gong, Y., Yang, Z., Chuai, G., Qu, S., Shen, A., Yu, J., & Liu, Q. (2019). Advances and challenges in deep generative models for de novo molecule generation. Wiley Interdisciplinary Reviews: Computational Molecular Science, 9(3). https://doi.org/10.1002/wcms.1395