GPT like transformer based conditional molecule generator and a high drug-likeness (QED) dataset generation

Wen Xing*

Department of sustainable energy, SINTEF industry, Norway

Juan Yang

Department of process chemistry and functional materials, SINTEF industry, Norway

June 7, 2024

Abstract

This study presents the development and evaluation of a novel GPT-like conditional molecule generator designed to optimize the synthesis of chemical compounds with desirable properties. The model incorporates six pivotal physicochemical properties as conditions: molecular weight, number of non-hydrogen atoms, ring count, hydrophobicity, quantitative estimation of drug-likeness (QED), and synthetic accessibility score (SAS). By integrating these specific attributes, the generator successfully produced a high-QED database, consisting of approximately 2 million molecules, all exhibiting a QED higher than 0.9. This achievement not only demonstrates the model's effectiveness in generating structurally diverse and potentially pharmacologically viable molecules but also underscores its utility in accelerating drug discovery processes.

GPT like, conditional generation, QED, molecules, transformer

*wen.xing@sintef.no

1 Introduction

In the rapidly evolving field of drug discovery and chemical synthesis, the ability to generate novel molecules with desired properties has become a critical paradigm. Leveraging the power of transformer-based models, like Generative Pre-trained Transformers (GPT), offers a promising approach to address this challenge[1]. Such models, once trained on a substantial dataset, can be fine-tuned to produce chemical structures that not only satisfy specific chemical criteria but also may possess desirable pharmacological attributes. In the burgeoning field of molecular design, the application of machine learning (ML) techniques has revolutionized the way scientists generate and optimize novel chemical entities. Various ML architectures, including RNN-based[2, 3, 4] , LSTM-based[5, 6, 7], Transformer-based[8, 9, 10], Variational Autoencoders (VAE)[11, 12, 13, 14], and Generative Adversarial Networks (GANs)[15, 16, 17], along with specialized models like conditional Generative Pre-trained Transformers (GPT), offer unique capabilities and challenges in the generation of molecular structures.

This study focus on a GPT-like transformer model, specifically designed for molecule generation, considering six pivotal physicochemical properties: molecular weight, number of non-hydrogen atoms, ring count, hydrophobicity, quantitative estimation of drug-likeness (QED)[18], and synthetic accessibility score (SAS). Developing a model that can efficiently manipulate these dimensions in molecular design can immensely benefit synthetic chemistry, reducing the time and cost associated with experimental methodologies. This conditional GPT-like model for molecule generation, specifically configured to incorporate six selected physicochemical properties as mentioned above. While these properties were chosen due to their significant relevance in molecules designing, it is important to underscore that the architecture of the model is inherently flexible. The conditional nature of this GPT-like generator is not limited to these six attributes alone; indeed, the framework can be readily adapted to integrate a wide array of other molecular properties and conditions. This versatility allows for customization to meet specific research needs and objectives, making it an invaluable tool in the exploration and design of novel compounds.

2 Methods

In our study, we employed a autoregessive attention only network[19], recognized for its robustness in modeling complex data relationships across fields such as natural language processing and cheminformatics. This architecture's effectiveness lies in its capacity to handle intricate dependencies and interactions, making it ideal for the nuanced task of generating molecules with specific characteristics. The core configuration of the model includes a hidden size of 512, which provides substantial capacity to understand complex data relationships while maintaining manageability on standard computational hardware. Additionally, the model features a multi-head attention mechanism with 16 heads, allowing simultaneous processing of diverse data facets, and is built with a depth of 24 layers to capture deep hierarchical patterns essential for accurate molecule synthesis. Despite its robustness, this model configuration is smaller compared to larger language models used in broader tasks, a design choice driven by the specific requirements of molecule generation. This smaller scale ensures the model remains both efficient and practical for the task at hand, balancing computational demands with performance. Compared to larger models, our transformer demonstrates a strong capability to predict molecular structures in response to set conditions such as molecular weight, atom count, and synthetic accessibility. The decision to use a relatively compact model configuration was informed by extensive hyperparameter tuning and existing research, which suggested diminishing returns for larger models in tasks of similar complexity. This strategic choice ensures that our model can be effectively trained and utilized with the available resources, making it well-suited for iterative development and real-world applications, thus providing a powerful yet computationally feasible tool for exploring complex relationships in molecule generation tasks.

In the development of our model for embedding SMILES (Simplified Molecular Input Line Entry System) strings, we opted for a character-level encoding approach to construct the vocabulary. This decision was driven by the simplicity and directness of treating each character in the SMILES notation as an independent token, which simplifies the model architecture and training process. Character-level tokenization ensures comprehensive coverage of the chemical space without the risk of missing rare or novel substructures that might not be included in a predefined word dictionary.

However, it is acknowledged that utilizing a group-based word dictionary, such as that employed in ChemBerta[20], could potentially enhance the

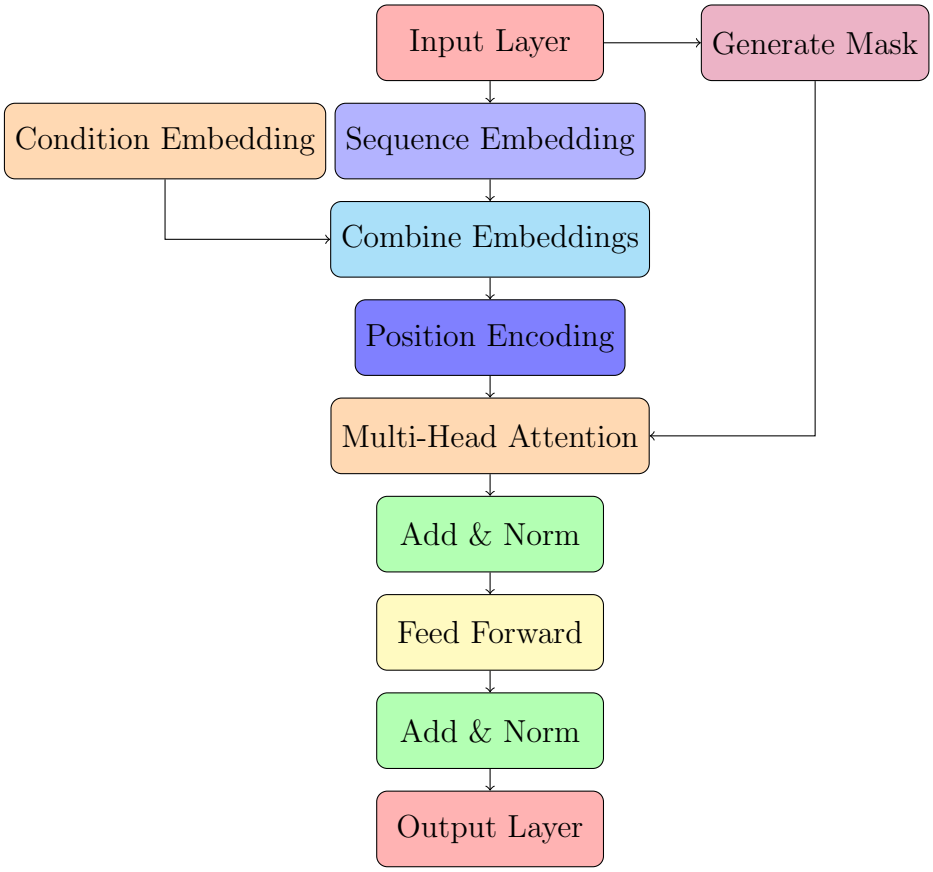


Figure 1: *Model architecture based on the vanilla attention only model*[19]*, the hidden-size was set to be 512, the number of heads was 16 and the layers was 24.*

model's performance, particularly in the accurate prediction of valid molecular structures. ChemBerta and similar approaches leverage groupings of characters that represent common chemical substructures, thus encapsulating more contextual information per token than individual characters. This method could reduce the complexity of the model and improve its learning efficiency by decreasing the sequence length of the inputs and capturing more meaningful chemical patterns.

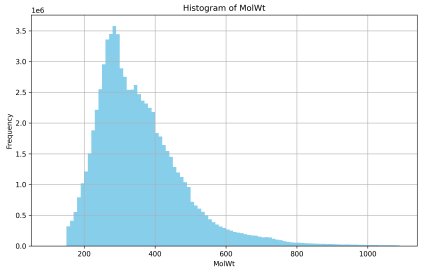
3 Model training

3.1 Dataset

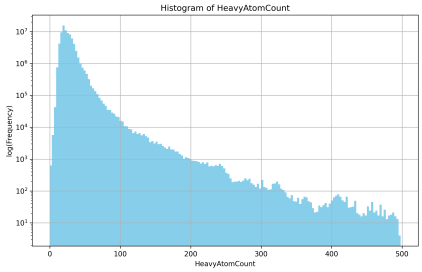
For training, we utilized an extensive dataset from PubChem[21], which included around 77 million molecules. To ensure manageability and relevance, only those molecules with a length of fewer than 200 characters were selected, aligning with typical lengths of small to medium-sized molecules. This preselection criteria helped in maintaining a focus on computationally feasible molecular sizes during model training and subsequent generative tasks. In this study, we utilized the open-source dataset from PubChem comprising approximately 77 million molecules, previously employed in the training of the ChemBERTa model. To facilitate our analysis, we calculated six critical molecular properties for each molecule using the RDKit cheminformatics software[22]. These properties include Molecular Weight (MolWt), Heavy Atom Count, Ring Count, Molecular LogP (MolLogP), Quantitative Estimate of Drug-likeness (QED), and synthetic accessibility score (sascore). Each of these properties was chosen for its relevance in assessing the drug-likeness and synthetic feasibility of molecules, crucial factors in drug discovery and design. The Molecular Weight provides insight into the size of the molecule, Heavy Atom Count gives the number of non-hydrogen atoms, Ring Count indicates the number of ring structures, MolLogP measures lipophilicity, QED offers a quantitative estimate of drug-likeness, and sascore assesses the ease of synthesis. The distribution of each of these molecular properties across the dataset is illustrated in Figure 2. This visualization helps in understanding the chemical space covered by the dataset and the typical profiles of molecules it contains. Analyzing these distributions is essential for setting appropriate conditions for molecule generation, ensuring that the generated molecules are representative of realistic, synthesizable, and potentially pharmacologically active compounds.

3.2 Batch, optimization and regularization

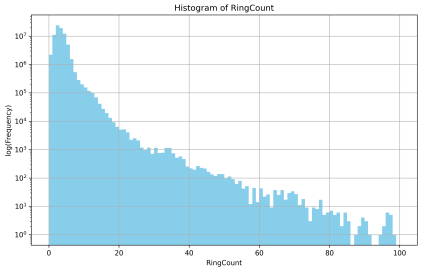
The training of the model was configured to accommodate the memory constraints of the available hardware, setting the batch size at 40. Optimization was performed using the Adam optimizer[23], with an initial learning rate of 1×10^{-5}. This learning rate was scheduled to decay by a factor of 0.2 starting from the second epoch to facilitate finer adjustments in the later stages of



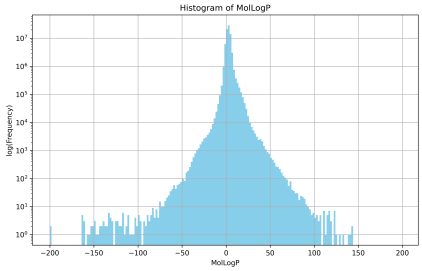
(a) molecule weight



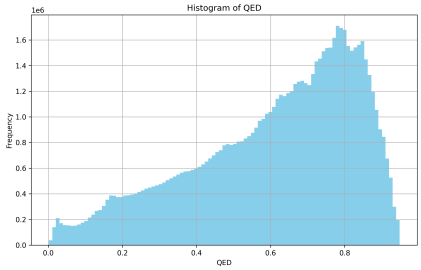
(b) non-hydrogen atom count



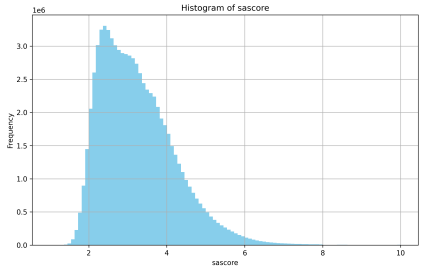
(c) Ring count



(d) Hydrophobic/hydrophilic



(e) Drug-likeness



(f) Synthesis score

Figure 2: *Distributions of the calculated properties for the training molecules data*

Smiles	MolWt	HeavyAtomCount	RingCount	MolLogP	QED	sascore
CCC=CCC=CC...	457.548	31	0	3.679	0.163	3.726
CC(C)(C)n1...	304.269	19	1	1.037	0.845	3.424
C[NH+](C)C...	455.604	32	3	2.703	0.386	2.949
Cc1cc(C)c(...	245.254	18	2	2.361	0.832	2.853
...	...	...	...	...	...	...

Table 1: *Data structure for the training dataset, total amount of the data is approximately 77 million molecules*

training. To further optimize resource utilization and improve computational efficiency, the model was trained using half-precision floating-point (FP16) arithmetic. However, this configuration occasionally led to instability in the training process, manifesting as non-converging numerical values (NaN) in the loss calculations. for each transformer decoder layer, we set a dropout ratio of 0.1.

3.3 Hardware

The model has been trained on a RTX4090 graphic card with 2 epochs. Using the hypermetropias described in section of Method, each epoch took about 17 hours to finish ending up of about 54 hours training time.

4 Testing the model

The efficacy of the model was assessed through its ability to generate a high-quality dataset characterized by a high Quantitative Estimate of Drug-likeness (QED). The methodology employed for this evaluation is detailed in the subsequent section.

Upon processing 7.2 million conditions through the generator, the model successfully produced 6.5 million valid SMILES strings, achieving a generation success rate of 90.27%. Among these valid SMILES strings, 4,636,508 were unique, representing a novelty rate of 71.33%. And compared with the original training data 4.43% of the generated molecules is the same in training database, representing a novelty of generation of 95.57%.

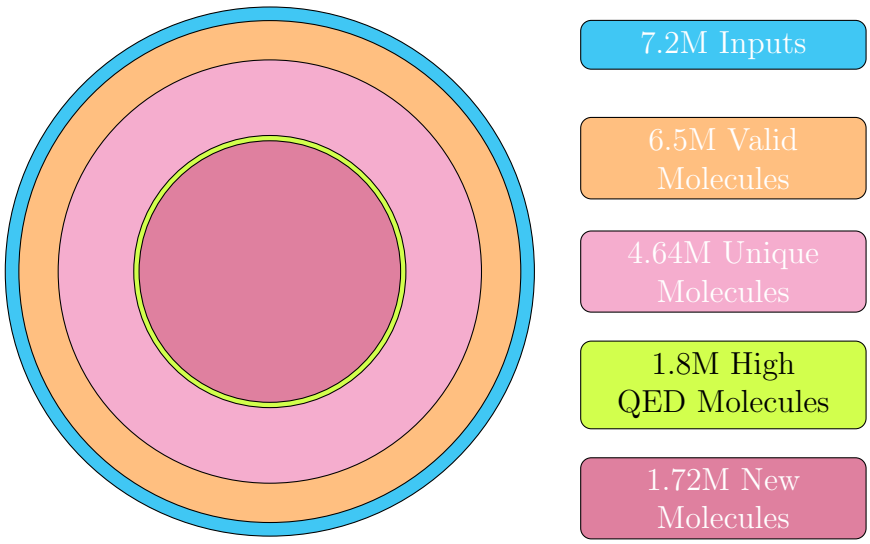


Figure 3: *Visualization of molecules generation validation and novelty*

5 Generation high QED molecules

In this research, we extend the methodology for generating novel molecular datasets by leveraging a pretrained Generative Pre-trained Transformer (GPT)-like model, incorporating specific conditions to guide the generation process. To evaluate the efficacy of this conditional generation approach, we initially sampled six million combinations from cross distributions of multiple molecular descriptors—Molecular Weight (MolWt), Heavy Atom Count, Ring Count, Molecular LogP (MolLogP), and synthetic accessibility score (sascore). These descriptors were selected from a comprehensive dataset comprising 77 million molecules, each characterized by a calculated Quantitative Estimate of Drug-likeness (QED) exceeding 0.85, emphasizing the focus on potentially high-quality drug-like molecules. For the generation phase, we employed these sampled descriptor values as input conditions, setting the QED condition explicitly to unity (1.0) to simulate an ideal scenario in drug design where maximum drug-likeness is desired. This approach allowed us to systematically assess how well the GPT-like model can adhere to and reproduce these optimal conditions in the generated molecules. By integrating these stringent conditions, we aim to demonstrate that our model can not only recreate molecules resembling those in the high-quality subset of the training data but also potentially innovate by exploring the chemical

space around these high-condition benchmarks. This strategy highlights the model's capability to generate viable molecular candidates tailored to specific pharmacological profiles, thereby enhancing the utility of generative models in drug discovery and development.

To ensure that the input conditions for our GPT model reflect realistic molecular properties, we first performed a comprehensive statistical analysis of our molecular dataset. We segmented molecular properties into bins based on their Quantitative Estimate of Drug-likeness (QED) scores and calculated the mean and covariance of critical molecular descriptors within each bin. These descriptors included molecular weight, heavy atom count, ring count, partition coefficient, QED itself, and synthetic accessibility score. The binning approach facilitated a nuanced understanding of property distributions across different levels of drug-likeness.

Utilizing the calculated means and covariance matrices, we developed a function to draw samples that conform to the observed empirical distributions of molecular properties. For each set of conditions—defined by a specific bin index and property column—this function generates samples from a multivariate normal distribution. The parameters of this distribution are directly derived from the mean and covariance statistics of the bin, ensuring that each sample reflects feasible combinations of molecular properties. This method significantly reduces the likelihood of proposing chemically implausible or unrealistic molecules as inputs to the generative model.

The sampled molecular properties serve as conditioned inputs to our GPT-based molecular generator. By feeding the model with inputs that are statistically representative of realistic molecular configurations, we effectively guide the generation process towards chemically viable and drug-like molecules. This approach not only enhances the chemical relevance of the generated molecules but also streamlines the discovery process by focusing on candidates with higher potential for successful synthesis and development.

6 Conclusion

The development of the GPT-like conditional molecule generator marks a significant advancement in the field of computational chemistry and drug design. The ability to generate a database of 2 million molecules with high quantitative estimation of drug-likeness (QED ¿ 0.9) validates the model's capability to effectively integrate critical physicochemical properties into the

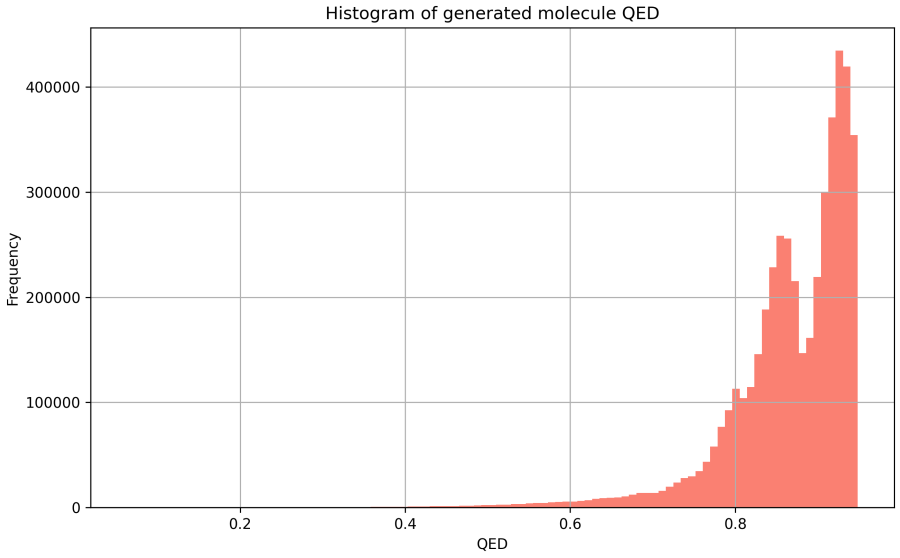


Figure 4: *Distributions of the QED for generated molecules*

generation process. This study's outcomes highlight the potential of machine learning models to contribute meaningfully to the early stages of drug discovery by providing a rapid, reliable means to generate compounds with targeted properties. Future work will focus on refining the model's accuracy and expanding its application to include a broader range of chemical entities, ultimately aiming to enhance the efficiency and innovativeness of drug development.

7 Code and dataset availability

The code for training the model can be found in Github (working on the repo...) the generated high QED dataset can be found at https://dx.doi.org/10.21227/egvm-m266 A demostration web application for this model can be found at https://buluway.com/mol

References

[1] Xin Qi et al. "Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges". In: *Molecules* 29.4 (Feb. 18, 2024), p. 903. ISSN: 1420-3049. DOI: 10.3390/molecules29040903. pmid: 38398653. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10892089/ (visited on 05/10/2024).

[2] Jinping Zou, Long Zhao, and Shaoping Shi. "Generation of Focused Drug Molecule Library Using Recurrent Neural Network". In: *Journal of Molecular Modeling* 29.12 (Nov. 6, 2023), p. 361. ISSN: 0948-5023. DOI: 10.1007/s00894-023-05772-5. URL: https://doi.org/10.1007/s00894-023-05772-5 (visited on 05/08/2024).

[3] Pengwei Hu et al. "De Novo Drug Design Based on Stack-RNN with Multi-Objective Reward-Weighted Sum and Reinforcement Learning". In: *Journal of Molecular Modeling* 29.4 (Mar. 30, 2023), p. 121. ISSN: 0948-5023. DOI: 10.1007/s00894-023-05523-6. URL: https://doi.org/10.1007/s00894-023-05523-6 (visited on 05/08/2024).

[4] Chuan Li et al. "Correlated RNN Framework to Quickly Generate Molecules with Desired Properties for Energetic Materials in the Low Data Regime". In: *Journal of Chemical Information and Modeling* 62.20 (Oct. 24, 2022), pp. 4873–4887. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.2c00997. URL: https://doi.org/10.1021/acs.jcim.2c00997 (visited on 05/08/2024).

[5] Maud Parrot et al. "Integrating Synthetic Accessibility with AI-based Generative Drug Design". In: *Journal of Cheminformatics* 15.1 (Sept. 19, 2023), p. 83. ISSN: 1758-2946. DOI: 10.1186/s13321-023-00742-8. URL: https://doi.org/10.1186/s13321-023-00742-8 (visited on 05/08/2024).

[6] Bongsung Bae, Haelee Bae, and Hojung Nam. "LOGICS: Learning Optimal Generative Distribution for Designing de Novo Chemical Structures". In: *Journal of Cheminformatics* 15.1 (Sept. 7, 2023), p. 77. ISSN: 1758-2946. DOI: 10.1186/s13321-023-00747-3. URL: https://doi.org/10.1186/s13321-023-00747-3 (visited on 05/08/2024).

[7] Marcos V. S. Santana and Floriano P. Silva-Jr. "De Novo Design and Bioactivity Prediction of SARS-CoV-2 Main Protease Inhibitors Using Recurrent Neural Network-Based Transfer Learning". In: *BMC Chemistry* 15.1 (Feb. 2, 2021), p. 8. ISSN: 2661-801X. DOI: 10.1186/s13065-021-00737-2. URL: https://doi.org/10.1186/s13065-021-00737-2 (visited on 05/08/2024).

[8] Changnan Gao et al. "DockingGA: Enhancing Targeted Molecule Generation Using Transformer Neural Network and Genetic Algorithm with Docking Simulation". In: *Briefings in Functional Genomics* (Apr. 6, 2024), elae011. ISSN: 2041-2657. DOI: 10.1093/bfgp/elae011. URL: https://doi.org/10.1093/bfgp/elae011 (visited on 05/08/2024).

[9] Yasuhiro Yoshikai et al. *A Novel Molecule Generative Model of VAE Combined with Transformer for Unseen Structure Generation*. Apr. 5, 2024. DOI: 10.48550/arXiv.2402.11950. arXiv: 2402.11950 [physics, q-bio]. URL: http://arxiv.org/abs/2402.11950 (visited on 05/08/2024). preprint.

[10] Jiashun Mao et al. "Transformer-Based Molecular Generative Model for Antiviral Drug Design". In: *Journal of Chemical Information and Modeling* 64.7 (Apr. 8, 2024), pp. 2733–2745. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.3c00536. URL: https://doi.org/10.1021/acs.jcim.3c00536 (visited on 05/08/2024).

[11] Wonho Zhung, Hyeongwoo Kim, and Woo Youn Kim. "3D Molecular Generative Framework for Interaction-Guided Drug Design". In: *Nature Communications* 15.1 (Mar. 27, 2024), p. 2688. ISSN: 2041-1723. DOI: 10.1038/s41467-024-47011-2. URL: https://www.nature.com/articles/s41467-024-47011-2 (visited on 05/08/2024).

[12] Divahar Sivanesan. *Attention Based Molecule Generation via Hierarchical Variational Autoencoder*. Jan. 18, 2024. DOI: 10.48550/arXiv.2402.16854. arXiv: 2402.16854 [cs, q-bio]. URL: http://arxiv.org/abs/2402.16854 (visited on 05/08/2024). preprint.

[13] Arun Singh Bhadwal, Kamal Kumar, and Neeraj Kumar. "NRC-VABS: Normalized Reparameterized Conditional Variational Autoencoder with Applied Beam Search in Latent Space for Drug Molecule Design". In: *Expert Systems with Applications* 240 (Apr. 15, 2024), p. 122396. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.122396. URL: https://www.

sciencedirect.com/science/article/pii/S0957417423028981 (visited on 05/08/2024).

[14] Hiroaki Iwata et al. "VGAE-MCTS: A New Molecular Generative Model Combining the Variational Graph Auto-Encoder and Monte Carlo Tree Search". In: *Journal of Chemical Information and Modeling* 63.23 (Dec. 11, 2023), pp. 7392–7400. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.3c01220. URL: https://doi.org/10.1021/acs.jcim.3c01220 (visited on 05/08/2024).

[15] Chen Li and Yoshihiro Yamanishi. "TenGAN: Pure Transformer Encoders Make an Efficient Discrete GAN for De Novo Molecular Generation". In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, Apr. 18, 2024, pp. 361–369. URL: https://proceedings.mlr.press/v238/li24d.html (visited on 05/08/2024).

[16] Xiaohong Liu et al. "MolFilterGAN: A Progressively Augmented Generative Adversarial Network for Triaging AI-designed Molecules". In: *Journal of Cheminformatics* 15.1 (Apr. 8, 2023), p. 42. ISSN: 1758-2946. DOI: 10.1186/s13321-023-00711-1. URL: https://doi.org/10.1186/s13321-023-00711-1 (visited on 05/08/2024).

[17] Zhiwen Zhu et al. "Automated Generation and Analysis of Molecular Images Using Generative Artificial Intelligence Models". In: *The Journal of Physical Chemistry Letters* 15.7 (Feb. 22, 2024), pp. 1985–1992. DOI: 10.1021/acs.jpclett.3c03504. URL: https://doi.org/10.1021/acs.jpclett.3c03504 (visited on 05/08/2024).

[18] G. Richard Bickerton et al. "Quantifying the Chemical Beauty of Drugs". In: *Nature Chemistry* 4.2 (Feb. 2012), pp. 90–98. ISSN: 1755-4349. DOI: 10.1038/nchem.1243. URL: https://www.nature.com/articles/nchem.1243 (visited on 05/10/2024).

[19] Ashish Vaswani et al. *Attention Is All You Need*. Aug. 1, 2023. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 05/08/2024). preprint.

[20] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. *ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction*. Oct. 23, 2020. DOI: 10.48550/arXiv.2010.09885. arXiv:

2010.09885 [physics, q-bio]. URL: http://arxiv.org/abs/2010.09885 (visited on 05/10/2024). preprint.

[21] Sunghwan Kim et al. "PubChem 2023 Update". In: *Nucleic Acids Research* 51.D1 (Jan. 6, 2023), pp. D1373–D1380. ISSN: 0305-1048. DOI: 10.1093/nar/gkac956. URL: https://doi.org/10.1093/nar/gkac956 (visited on 05/08/2024).

[22] *RDKit*. URL: https://www.rdkit.org/ (visited on 05/10/2024).

[23] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. Jan. 29, 2017. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 [cs]. URL: http://arxiv.org/abs/1412.6980 (visited on 05/10/2024). preprint.