Using Transformers for De Novo Drug Molecule Generation

PROJECT REPORT

*submitted in partial fulfilment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

*in*

**CHEMICAL ENGINEERING**

By

Ujjwal Anand (2001CB60)

*Under the supervision of*

**Dr. Atanu K. Metya**



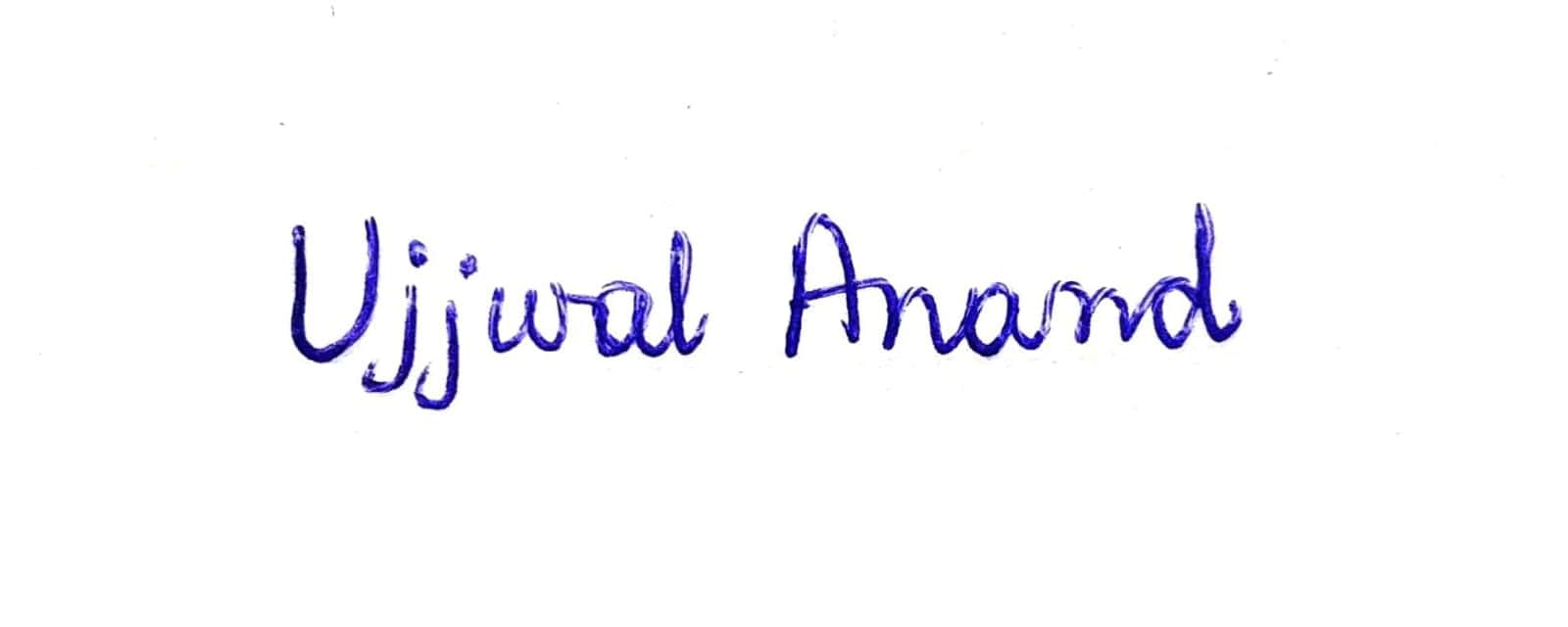
Department of Chemical and Biochemical Engineering Indian Institute of Technology Patna

**May 2024**

DECLARATION

I hereby declare that this report is my original work, and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the report. I have not plagiarized any content of this report.

This report has not been submitted in any other institute or university for the award of any degree or diploma.



|  |  |
| --- | --- |
| IIT Patna  May 2024 | **Ujjwal Anand (2001CB60)** |

CERTIFICATE

This is to certify that the thesis titled “**Using Transformers for De Novo Drug Molecule Generation**’’ submitted by **Mr. Ujjwal Anand** to Indian Institute of Technology Patna for the award of the degree of Bachelor of Technology, has been carried out under my supervision. The same has not been submitted elsewhere for a degree.

Supervisor: **Dr. Atanu K. Metya**

Assistant Professor,

Department of Chemical and Biochemical Engineering,

Indian Institute of Technology Patna

ABSTRACT

This work investigates the applicability of transformer designs in drug design, with a particular emphasis on the synthesis of similar molecules from known pharmacological compounds. Drug optimization is a large and discontinuous domain with many issues that deep learning techniques, in particular, decoder transformers, try to solve.

Many features of the model have been explained, including self-attention, its potential for use in comprehending intricate interatomic interactions. The advantages of using transformers over more established deep learning techniques like RNNs, Autoencoders, and GANs, for de novo drug design has also been discussed.

Acknowledgment

In submitting this thesis to the Department of Chemical and Biochemical Engineering, Indian Institute of Technology Patna, in partial fulfilment of the requirements for the degree of Bachelor of Technology in Chemical Engineering, I affirm my awareness of the standards of the Ethical Code of Conduct of the Institute. It is the product of my labour, except wherever indicated in the text. The report may be freely copied and distributed, provided the source is acknowledged. I am writing to express my sincere gratitude to my mentor, **Dr. Atanu K. Metya,** for allowing me to work under his supervision. His valuable guidance and encouragement were vital, without which this project would not have come forth.

I would also like to convey my gratitude to my guide, Mr. Ganesh K. Reddy for assisting me throughout my journey.

Contents

Declaration

Certificate

Abstract

Acknowledgment

1. Introduction
2. Motivation for Project
3. Objective

4. Literature Review

5. Methodology

* + 1. Data Preparation
    2. Model building
    3. Training

6. Result and discussion

7. Conclusion

8. References

1. **Introduction**

Navigating the vast expanse of the optimization domain for drugs, coupled with its discontinuous nature, poses a significant challenge. Fortunately, deep learning methods offer an effective means to autonomously acquire task-relevant feature extraction.

The estimated count of drug-like molecules ranges from 1023 to 1080, depending on various estimation methods. Despite the rapid advancement in computing power, the exploration of such a vast number of molecules still demands considerable time and financial investment to unearth potential candidate compounds. [1]

De novo drug design, a conventional approach, entails crafting fresh ligands within receptor pockets through the amalgamation of fragments or atoms, leveraging established algorithms like genetic and Monte Carlo methods [2]. Conversely, machine learning, offering a distinct avenue from conventional drug design, has significantly influenced de novo drug design. Traditional machine learning techniques manually extract data features tailored to specific domains, thus constraining their capacity to handle raw data comprehensively. However, with the advent of advanced machine learning and big data technologies, more robust methodologies, such as deep learning, have emerged. These methodologies are capable of discerning concealed functional correlations within original datasets and subsequently applying them to novel data instances. Deep learning employs layers of nonlinear transformations to extract and amalgamate information from data, facilitating the discernment of broader trends and patterns amidst noise. Furthermore, it presents the prospect of uncovering features yet to be identified by researchers, thereby transcending the confines of current knowledge. Specifically, deep learning techniques are exceptionally well-suited for de novo drug design. Hence, several studies demonstrate their efficacy in enhancing drug development by identifying more favourable compound candidates and augmenting the success rate of clinical trials for these candidates.

2. Motivation for the project

De novo drug design involves constructing fresh ligands within designated receptor pockets through the assembly of fragments or atoms. Deep learning methods present an alternative avenue for de novo drug design by leveraging the training of chemical and physical properties inherent in drugs. The pivotal stage in de novo drug design via deep learning methods lies in extracting the requisite chemical and physical features.

In the realm of deep learning-based drug discovery, SMILES strings, representing specific chemical structures, serve as popular inputs/outputs. Recent years have witnessed numerous endeavours aimed at facilitating the conditional generation of drug-like molecules, depicted as SMILES strings, with predefined properties. Notably, recent studies have demonstrated the capability of variational autoencoders to generate molecules with specific properties by leveraging concatenated SMILES strings containing the desired property information.

Moreover, Recurrent Neural Network (RNN)-based generative models have undergone extensive evaluation in molecular design. An alternative approach involves modifying an existing SMILES string by introducing chemically informed alterations. Subsequently, both Transformer and RNN seq2seq architectures are tested by feeding this modified SMILES string as input and generating a SMILES string with the desired properties as output.

In drug design employing deep learning methodologies, encoder-decoder architectures such as autoencoders are frequently utilized. These architectures are trained iteratively to minimize the disparity between the reconstructed output and the original input. The primary objective of the autoencoder is to uncover a more compact representation of samples.

Thus, they compress the incoming input information to a latent vector space representation. They then sample continuously from the latent space in the region around the original drug molecule to get various new configurations of the drug. This causes a loss of the contextual information stored in each of the atoms and the model does not learn the proper way that atoms should be arranged, and the output configurations are mostly invalid and infeasible to be produced.

Similarly, generative adversarial networks comprise of a generator G and a discriminator D. Typically, the generator is tasked with learning to map random noise to a specific distribution closely resembling the data distribution, while the discriminator discerns whether the input is genuine data or a sample generated by the generator, functioning essentially as a binary classifier. Once the model is adequately trained, new samples can be generated from the generator.

In the adversarial process, both the generator ‘G’ and discriminator ‘D’ are simultaneously trained. ‘D’ endeavours to uncover hidden patterns in the input data, accurately distinguishing real data from data generated by ‘G’, while ‘G’ refines its weights through iterative optimization of data sampling, aiming to outsmart the well-trained ‘D’. [2]

However, the objective of a GAN is not to learn the relationships between individual atoms, and they just replicate a new molecule as close as possible to the original drug molecule. Thus, they do not capture inter-atomic relationships and dependencies well and suffer from the same problems of invalid and infeasible molecular outputs, as the autoencoder architecture.

In our study, we adopt a Natural Language Processing (NLP) approach and perceive the small-molecule drug design conundrum as a text/SMILES generation challenge. Various deep learning techniques have proven successful in text generation. For instance, the GPT (Generative Pre-Trained Transformer) series utilizes an autoregressive language model to generate human-like text, having been trained on extensive amounts of unlabelled human-written text. The text generated by the GPT model exhibits high quality and is challenging to differentiate from human-authored content. Similarly, GPT models demonstrate proficiency in learning chemical structures from vast molecular datasets. The fundamental concept of GPT involves acquiring natural language skills by predicting the subsequent word based on preceding words from a large text corpus through unsupervised learning. This unsupervised pre-training imbues the model with drug-related knowledge and ensures the generation of valid SMILES strings. A well-trained GPT model can produce synthetic text excerpts while conditioning on arbitrary inputs. Furthermore, the GPT architecture supports conditional generation through fine-tuning with small-sized supervised data.

**3. Objective**

The objective of this study is to use transformers for the task of drug-design which have been remarkably successful in the domain of Natural Language Processing recently. Transformers capture and encode relationships between individual tokens in a sentence through self-attention mechanisms. In a transformer model, each token in a sequence is associated with three learnable vectors: a query vector, a key vector, and a value vector. These vectors are derived from the embedding of the tokens in the input sequence.

This process is useful because it allows the transformer model to address relevant tokens in the input sequence and capture complex relationships between them. By encoding these relationships, the model can better understand the context of each token and produce more accurate molecules with significantly higher validity. Additionally, transformers are highly parallelizable, making them efficient for processing large datasets and achieving state-of-the-art performance.

**4. Literature review:**

To learn about various architectures used by researchers, a systematic search was conducted across various academic databases and the code was replicated to understand its workings. The review focused on encoder-decoder architectures, generative adversarial networks, and transformers.

Gomez-Bombarelli et al. [3] were the first to utilize a VAE framework in chemical design, introducing a method termed ChemicalVAE. Within this autoencoder model, molecules represented as SMILES strings were transformed into a latent space, capturing distinctive features of the training data. Although the autoencoder could undergo joint training with a property prediction task to enhance molecular properties, a notable issue with this model was the occasional generation of invalid molecules or those containing undesirable moieties like acid chlorides or cyclo-butadienes.

In a similar vein, Kusner et al. [4] integrated context-free grammar into the VAE framework, imparting explicit knowledge on generating valid molecules. Their innovation, known as the grammar variational autoencoder (GrammarVAE), exhibited a higher percentage of valid molecules by generating syntactically correct SMILES strings and achieving a smoother latent space. Nevertheless, certain non-context-free aspects of SMILES strings remained unaccounted for in this approach.

GANs have found application in molecular generation using SMILES representations. One of the first successes in this domain was demonstrated by Guimaraes et al. [5] who introduced ORGAN, a GAN framework integrated with Reinforcement Learning (RL) inspired by SeqGAN. The molecules generated by ORGAN exhibited enhanced diversity compared to previous approaches. In summary, these models have the capability to generate molecules that accurately represent the original data distribution, display improvements in desired metrics, and retain diversity among the generated samples.

Kadurin et al. [6] were the first to suggest employing Adversarial Autoencoders (AAEs) for the generation of novel compounds intended for cancer treatment. Their approach involved the conversion of SMILES strings obtained from PubChem into chemical fingerprints. This model was trained to encode and reconstruct not only molecular fingerprints but also experimental concentration data. Subsequently, the model was utilized to screen a vast dataset of 72 million compounds available on PubChem to identify potential candidate molecules with anticancer properties. The screening yielded 69 compounds spanning various chemical classes, among which it was discovered that some had previously been employed as anticancer agents.

Jiashun et al. [7] introduced a pioneering data-driven self-supervised pretraining generative model named "TransAntivirus" aimed at performing select-and-replace edits to transform organic molecules into desired properties suitable for the design of antiviral candidates. In contrast to many molecule generation models reliant on SMILES representations, they adopted an IUPAC-directed expansion of the SMILES molecule generation space. The findings revealed that TransAntivirus surpassed control models significantly across metrics such as novelty, validity, uniqueness, and diversity.

1. **Methodology**
2. **Dataset preparation**

To simplify computational tasks, encoded strings were constrained to a maximum length of 100 characters. SMILES (Simplified Molecular Input Line-Entry System) representations for 2.3 million compounds with SMILES strings of 100 characters or fewer were downloaded from the PubChem database. Right after the filtration, 3 million molecules were retained, and the dataset was partitioned into two sets—training and validation—using an 85:15 ratio. SMILES strings were encoded as text using a 45-character set. If a string was shorter than 100 characters, it was padded with spaces to reach the specified length. To ensure consistency and avoid dealing with multiple equivalent representations of the same molecule, only "canonicalized" forms of SMILES—a standardized version—were used for training.

1. **Model building**

The transformer architecture was used to generate analogous molecules from an existing drug molecule. In our case, the transformer was an encoder-decoder transformer that is primarily used for sequence-to-sequence tasks. At its core, the transformer architecture employs self-attention mechanisms to weigh the importance of different input elements (tokens) in a sequence when generating an output sequence. By considering the entire input sequence simultaneously, rather than processing it sequentially like recurrent neural networks (RNNs), transformers can capture relationships between distant tokens more effectively.

Figure 1 shows the components of the model [8], which are further explained below:

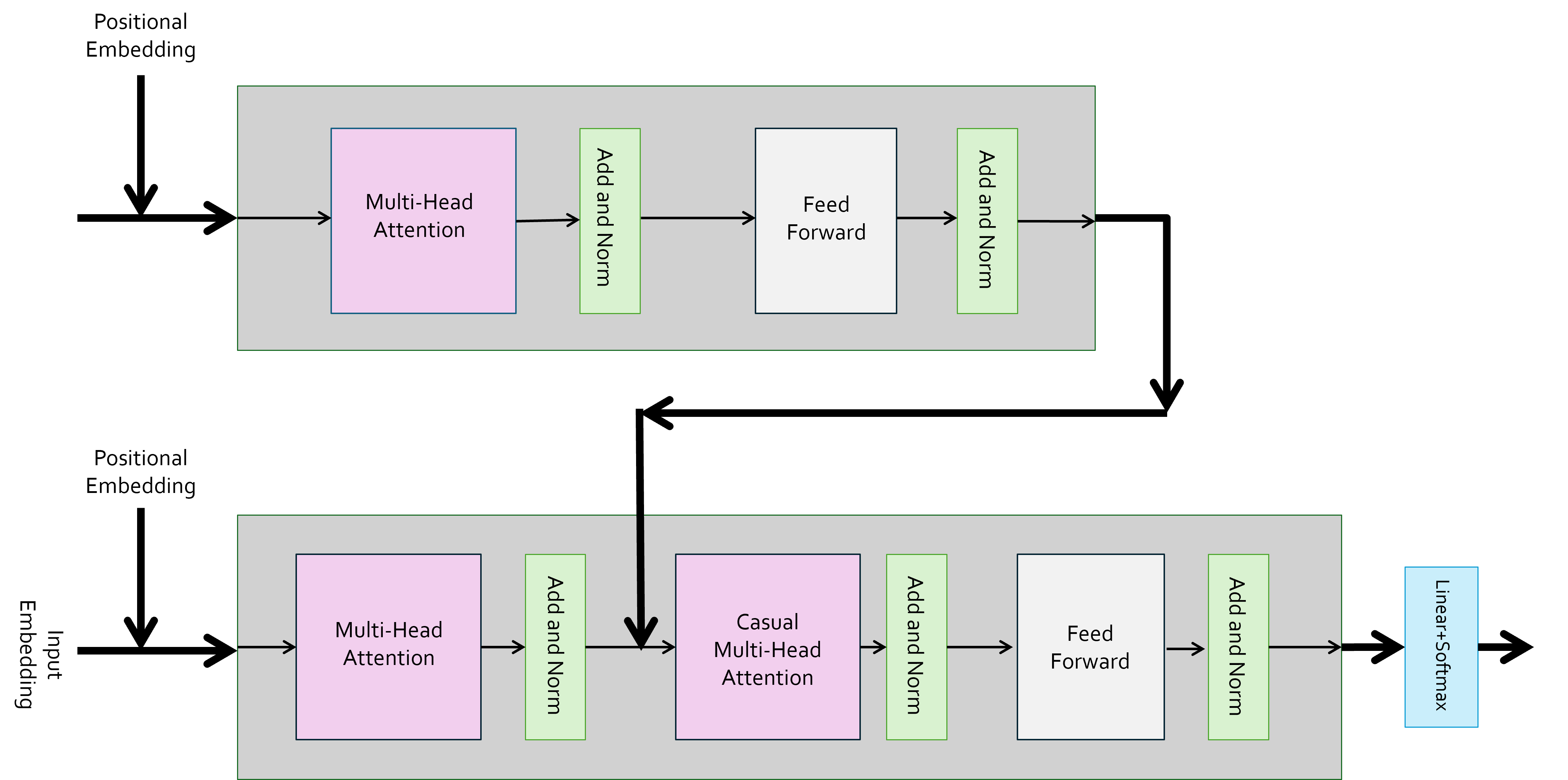


Fig 1: Architecture of a transformer consisting of an encoder conjoined with a decoder block, consisting of multi-head self-attention as well as cross self-attention.

**Input Embedding:** At the beginning of the model, input tokens (e.g., words or characters) are represented as high-dimensional vectors known as embeddings.

Each token is mapped to a continuous vector space where tokens with similar meanings or roles are closer together in the embedding space.

Input embeddings capture semantic information about the tokens and serve as the initial input.

**Positional Encoding:** To provide information about the position of tokens in the input sequence, positional encodings are added to the input embeddings. This allows the model to differentiate between tokens based on their position in the sequence.

**Masked Multi-Head Attention:** The masked multi-head attention layer is a crucial component of the decoder. In a traditional attention mechanism, there is a single attention head that computes attention scores between each token in the input sequence. This single attention head captures global dependencies in the input sequence but may struggle to capture more nuanced relationships between tokens.

To address the limitations of a single attention head, multi-head attention computes attention scores using multiple attention heads in parallel. For this purpose, the embedding for the input is fractionated row-wise, into several parts, and each part is called a ‘head’. Each attention head learns distinct aspects of the input sequence, allowing the model to capture diverse patterns and relationships.

In this layer, attention is computed across the decoder's input sequence, but with a mask applied to prevent attending to future tokens. This ensures that the model only has access to information from tokens that have already been generated.

In addition, multi-head attention is typically applied in multiple layers of the model. Each layer consists of multiple attention heads, and the outputs of the attention heads are concatenated and processed further through feedforward neural networks before being passed to the next layer.

**Encoder-Decoder Attention:**

Encoder-decoder attention is used to facilitate information flow between the encoder and decoder components. The encoder processes the input sequence and generates a sequence of context-aware representations. The decoder then utilizes these representations, along with its own self-attention mechanism, to generate the output sequence.

**Feedforward Networks**: Each decoder layer contains feedforward neural networks, which are responsible for processing the information obtained from the self-attention mechanism. These feedforward networks consist of two linear transformations with a non-linear activation function (in our case, ReLU) applied in between.

**Residual connections**: Also known as skip connections, these connections bypass one or more layers in a neural network and are added back to the output of subsequent layers. In the context of our architecture, residual connections can be added around each sub-layer, including the self-attention layer, feedforward network, and normalization layer. These connections allow information to flow more easily through the network and help mitigate the vanishing gradient problem, which can occur in deep neural networks.

By incorporating residual connections, the decoder-only architecture can facilitate the training of deeper networks and improve the overall performance of the model.

**Normalization:** For optimization of the input of each layer a pre-normalization layer is applied before each sub-layer in the decoder. Normalization helps stabilize training and improve the flow of gradients through the network.

**Dropout:** During training, dropout randomly sets a fraction of the neurons in a layer to zero, effectively "dropping out" those neurons temporarily. By randomly dropping out neurons, dropout prevents the network from relying too heavily on any single neuron or combination of neurons. This forces the network to learn more robust and redundant representations of the data.

**Output Projection:** At the output of the decoder layers, a linear projection layer is applied to transform the decoder's hidden representations into output logits. These logits represent the probabilities of each token in the output vocabulary, allowing the model to generate the next token in the sequence.

**Beam Search or Sampling:** Once the decoder produces the logits for the next token, a decoding strategy such as beam search or sampling is used to select the most likely token or generate multiple candidate sequences.

1. **Training of the model**

The objective of training the proposed deep learning model with the provided data was to minimize the dissimilarity between the input sequence and the reconstructed output sequence.

Using various batch sizes of 30, 50, 100, 300 it was found that the computation for large batch sizes was significantly slower compared to lower batch sizes, due to which a batch size of 30 was selected. Similarly, a greater embedding size means that more context can be integrated into the molecule, and so an embedding size of 512 was used.

The number of heads for multi-head attention was set to eight and three multi-head attention layers were used, along with a dropout percentage of 20%. The linear projection layers used 512 neurons compared to the fully connected layer with 2048 neurons.

The training process was run for 5 epochs and achieved a cross entropy loss of 0.09. During the training phase, the decoder aims to accurately reproduce the input chemical SMILES. Once the transformer has been trained, it can be utilized to generate analogous SMILES, also known as analogues, by encoding the corresponding chemical SMILES and employing the decoder to reconstruct the new SMILES. The resulting SMILES will resemble the original input, albeit with some modifications or variations induced by the noise present in the embedding space.

Operating in an autoregressive manner, the decoder part of the transformer generates one token at a time, with each token being dependent on the preceding ones. This sequential generation process allows for the systematic creation of output sequences. Utilizing self-attention, the model captures the interrelations and dependencies among the atoms within the molecule, resulting in more refined outputs compared to alternative architectures.

**6**. **Results and Discussion**

To evaluate the efficiency of this model over existing models, the same metrics from generated molecules and the original molecules were calculated using the MOSES (Molecular Sets) platform.

The model was compared to two models: an LSTM and a GAN architecture based on the parameters of Validity, Uniqueness and Novelty.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Validity** | **Uniqueness @1k** | **Uniqueness @10k** | **Novelty** |
| LSTM | 97.03% | 98.22% | 99.56% | 98.57% |
| GAN | 97.15% | 97.31% | 99.27% | 98.07% |
| Current model | 97. 55% | 96.25% | 98.85% | 98.45% |

In Table 1, we observe that our model performs better than LSTM and GAN in terms of validity and novelty of the generated molecules, and performs excellently in terms of uniqueness as well, only behind the LSTM architecture.

This model was also compared with a variational autoencoder (VAE) trained on the same dataset. For this, two important drugs Remdesivir and Bedaquiline were used, and the outputs obtained are as shown:

Table 1: Comparison of several architectures with the current model against parameters like Validity, Uniqueness@k, and Novelty

|  |  |  |
| --- | --- | --- |
| A structure of a molecule  Description automatically generated | A molecule structure with blue and black lines  Description automatically generated |  |
| A black and red molecule  Description automatically generated | A black and red molecule  Description automatically generated |
| Bedaquiline | From the current architecture | From the previous architecture |
|  | A molecule of a chemical structure  Description automatically generated with medium confidence |  |
| Remdesivir | From the current architecture | From the previous architecture |

Fig 2: Comparison of analogues generated for the drugs Remdesivir and Bedaquiline using the current transformer model and a variational autoencoder.

Here we observe that the analogues generated from our model are more valid and more feasible than those generated from a VAE. This is mainly attributed to the fact that the encoder-decoder transformers do not require a predefined latent space like VAEs do. This makes them more flexible and adaptable to several types of data and tasks. They can generate diverse outputs without being constrained by a fixed latent space structure, allowing for greater creativity and exploration in sequence generation.

In addition to the advantages offered by transformer architectures, some drawbacks were also observed, which are inherent to transformers. The model also showed a very low value for loss indicating overfitting in the network. To eliminate this, the size of the training dataset needs to be increased. More diverse and representative data can help the model generalize better to unseen examples, reducing the risk of overfitting. Early stopping may also be used to monitor the model's performance on a separate validation set during training and stopping the training process when the validation loss stops improving or begins to deteriorate.

This prevents the model from continuing to learn patterns specific to the training data that may not generalize well to unseen examples.

Another method proposed is by making the architecture of the transformer network deeper by increasing the number of layers, the dimensionality of the embeddings, or the number of attention heads can help prevent overfitting, especially if the model is excessively complex relative to the size of the training data.

Also, there is no particular method of expressing the accuracy of Generative models, and thus most of the inference is purely based on observations.

**7. Conclusion:**

In conclusion, this study demonstrates the effectiveness of decoder-only transformer architectures in drug design tasks, particularly in the generation of analogous molecules from existing drug compounds. By leveraging self-attention mechanisms and autoregressive generation processes, decoder transformers capture complex relationships within molecules and systematically generate output sequences.

In our investigation, the proposed framework was compared against other baseline models, revealing its superior performance in terms of novelty and validity compared to the control methods. Nonetheless, a limitation arises as the generated compounds fail to surpass the reference molecule, indicating a shortfall in uniqueness.

Comparative analyses with traditional models such as VAEs and LSTM RNNs highlight the superiority of decoder transformers in terms of computational efficiency and output quality. The findings of this study underscore the potential of decoder transformers as a valuable tool in drug design research, offering new avenues for the discovery and optimization of drug candidates. Future research may focus on further optimizing transformer architectures by reducing overfitting and training on larger datasets and exploring their applications in other domains of molecular design and drug discovery.

The model was implemented and shared in a Google Colaboratory notebook on the T4 GPU. The code and dataset for this project can be found on:

https://github.com/UjjwalAnand364/Drug-Lead-Optimization

**8. References**

1. Chakraborty, R; Hasija, Y. Utilizing deep learning to explore chemical space for drug lead optimization. Expert Systems with Applications, 229, 2023, 120592
2. Tong, X.; Liu, X.; Tan, X.; Li, X.; Jiang, J.; Xiong, Z.; Xu, T.; Jiang, H.; Qiao, N.; Zheng, M. Generative Models for De Novo Drug Design. Journal of Medicinal Chemistry, 2021, 64, 14011-14027
3. Gomez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; HernandezLobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; AguileraIparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using A Data-Driven Continuous Representation of Molecules. ACS Central Science, 2018, 4, 268−276
4. Kusner, M.; Paige, B.; Hernández-Lobato, J. Grammar Variational Autoencoder. In Proceedings of the 34th International Conference on Machine Learning, 2017, 70, 1945−1954
5. Guimaraes, G.L., Sánchez-Lengeling, B., Farias, P.L., & Aspuru-Guzik, A. (2017). Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. 2017, ArXiv, abs/1705.10843.
6. Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for De Novo Generation of New Molecules with Desired Molecular Properties in Silico. Molecular Pharmaceutics, 2017, 14, 3098−3104
7. Mao, J; Wang, J; Zeb, A; Kwang-Hwi Cho, Jin, H; J. Kim, Lee, O; Wang, Y; and Tai, K;. Transformer-Based Molecular Generative Model for Antiviral Drug Design. Journal of Chemical Information and Modeling, 2024, 64, 2733-2745
8. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; & Polosukhin, I. Attention is All you Need. Neural Information Processing Systems, 2017