university of groningen / faculty of mathematics and natural sciences / artificial intelligence

Bidirectional SELFIES for Molecular Design

Graduation Project Proposal
(Computational Intelligence and Robotics)

Andrei Voinea (s3754243)

April 3, 2023

Internal Supervisor: Dr. Matthia Sabatelli (Artificial Intelligence, University of Groningen)
External Supervisor: Dr. Robert Pollice (Chemistry, University of Groningen)

Artificial Intelligence
University of Groningen, The Netherlands

1 Introduction

The process of *de novo* molecular design involves proposing novel chemical structures that best meet a desired molecular profile. In the context of drug discovery, the aim is typically to elicit a desired biological response while also meeting acceptable pharmacokinetic criteria [16]. However, conventional approaches to molecular design are often hindered by the large number of potential molecules, estimated to be on the order of 10^{60} - 10^{100} [18]. This limitation results in a time-consuming and expensive trial-and-error process to synthesize and characterize molecules, which can take years to produce viable compounds. Moreover, as [18] argue, this process involves a wide range of parameters (e.g., structural features, additional constraints), which makes designing and optimizing molecules a challenging and multidimensional problem. Furthermore, this approach can be highly subjective and reliant on the intuition of experienced chemists, which lead to biases such as favouring known molecular structures or disregarding unconventional or innovative designs.

To overcome these challenges, researchers have turned to machine learning approaches to accelerate the molecular design process. By training models on large datasets of known molecules and their properties, machine learning algorithms can predict the properties of new molecules and suggest ones that may have desirable characteristics. These algorithms can also be used to optimize existing molecules to improve their properties, or to design novel molecules that are tailored to specific applications. Machine learning approaches have already shown promise in drug discovery, where they have been used to identify viable drug candidates [10, 9].

One particular area of machine learning for molecular design is the development of chemical language models, which use sequential string representations of molecules to generate new structures. One widely used representation is the Simplified Molecular-Input Line-Entry System (SMILES; 21), where a molecule is represented as a linear string of characters that encode its atoms, bonds, and functional groups. One advantage of SMILES is its compatibility with existing chemical databases and software tools, and it has shown promise in generating new molecules [1]. However, as the SMILES representation does not enforce explicit rules for the generation of valid molecular structures, it is prone to produce void strings. Additionally, the sequential nature of the SMILES representation restricts the possible modifications that can be made to a given molecule, and the resulting structures often share similar features. To address some of these limitations, previous research has proposed the use of strictly robust molecular string representations, such as the Self-Referencing Embedded Strings (SELFIES; 13).

The representation of SELFIES ensures that any arbitrary combination of allowed characters corresponds to a valid structure, thereby addressing a critical challenge in the generation of molecular structures. However, small random modifications of SELFIES may result in structures with limited diversity, which can limit exploration. Therefore, recent research has focused on the development of efficient chemical language models to generate diverse and realistic molecular structures, addressing the limitations of both SMILES and SELFIES. One such approach involves bidirectional generation of molecules, which has demonstrated remarkable outcomes while utilizing the SMILES representation [9].

To this end, the goal of this project is to evaluate the performance of using the SMILES and SELFIES representation across various language model architectures, and to identify steps to enhance existing methodologies. In the following sections, we will delve into further details regarding language models, and how these can be used to achieve *de novo* molecule generation.

2 Theoretical Framework

To understand the relevance of using transformer models in the context of *de novo* molecule generation, this section further describes the two string representations of molecules, SMILES and SELFIES. Furthermore, we briefly describe the attention mechanism of transformer models, and how it can be used to achieve a behaviour similar to previous bidirectional approaches.

2.1 String representations of molecules

As described previously, SMILES is a widely used notation for representing the structure of a molecule using a simple string of characters. The SMILES notation is based on a graph representation of the molecule, where atoms are represented as nodes, and bonds as edges. According to [21], to obtain a SMILES notation, the target molecule has to be parsed based on a set of rules. Firstly, atoms are denoted by their atomic symbols, and non-hydrogen atoms have their symbols enclosed in square brackets. Two-character symbols have the second letter in lower case (e.g., [C] for carbon and [Fe] for iron). Bonds can be single (—), double (=), triple (#), or aromatic (:). Parentheses indicate branches, and rings are indicated by breaking one single (or aromatic) bond and appending a digit after the atom at each ring closure.

Given these rules, it is possible for a molecule to be represented by multiple SMILES strings, which can hinder the process of generating unique molecules. Similarly, by introducing small changes in a SMILES representation, the resulting molecule can become invalid. To avoid the latter issue, [13] propose a formal Chomsky type-2 grammar, which uses a set of derivation rules to determine how to add atoms, bonds, or branches to a string representation. The derivation starts with a state X_0, and iteratively appends symbols until a terminal symbol is met. SELFIES allows the use of simple operations to update a molecule representation, such as insertion, deletion, or substitution of characters. Furthermore, since the derivation rules ensure syntactical and chemical constraints (e.g., the maximal number of valence bonds, ring closures, and branchings), SELFIES are exceptionally suited for machine learning tasks. As indicated by [13], models using this representation always generate valid molecules, and are capable of encoding more structural information during training.

2.2 Bidirectional approaches

As SELFIES are a recent development, a considerable number of studies have concentrated on employing SMILES to create molecules [20, 9, 10]. One notable approach to molecule generation has been proposed by [9], which utilizes a bidirectional recurrent neural network (RNN) to generate SMILES strings. In this approach, the RNN reads the SMILES sequence in both forward and backward directions, allowing it to capture dependencies and patterns that may exist in either direction. At each time step, the RNN predicts the next token in the SMILES string based on the current hidden state and the previous token in the sequence. The algorithm generates a new token in either the forward or backward direction, depending on the time step.

The bidirectional RNN architecture used in this approach allows the model to consider both past and future context when generating each character in the SMILES string. This can be particularly useful for generating longer and more complex SMILES strings, as the model can use information from both the

beginning and end of the string to inform its predictions. However, the use of RNNs can be computationally expensive and difficult to train, especially for longer sequences. Additionally, RNNs can suffer from vanishing or exploding gradients, which can limit the model's ability to learn complex patterns and dependencies in the data.

2.3 The transformer architecture

Recently, transformer-based models have emerged as a promising alternative for molecule generation. These models use attention mechanisms to learn dependencies between all positions in the input sequence, rather than relying on a fixed, sequential order. This allows transformer models to handle longer sequences and capture more complex patterns in the data. Transformer-based models have been shown to outperform RNN-based models on several molecule generation tasks, including generating diverse and novel molecules with high validity and similarity to known molecules [20].

The transformer architecture relies on the attention mechanism, which allows the model to selectively focus on different parts of the input sequence when making predictions [19]. The attention mechanism calculates a set of weights, or attention scores, that indicate how much each position in the input sequence should contribute to the output at each position in the output sequence. The transformer architecture consists of an encoder and a decoder, both of which use multiple layers of self-attention and feed-forward neural networks. The encoder processes the input sequence and produces a set of hidden representations, while the decoder generates the output sequence based on these hidden representations. This self-attention mechanism allows the model to learn to focus on different parts of the input sequence at different positions in the output sequence, which can be useful for modelling dependencies between different parts of the input and output sequences. Additionally, the parallel nature of the transformer architecture allows for faster training and inference compared to sequential RNN-based models.

Due to the self-attention mechanism, certain transformer-based language models already exhibit a bidirectional behaviour similar to the one proposed by [9]. One example is the Bidirectional Encoder Representations from Transformers (BERT; 4) model, which is pre-trained using a masked language modelling (MLM) task and a next sentence prediction (NSP) task. During the MLM task, a set number of tokens is replaced with a "mask" token, which the model has to predict in order to recreate the original input. On the other hand, during the NSP task, the model has to predict whether two given sequences are consecutive or not. These two tasks ensure that the model learns to capture both local and global dependencies in the input sequence.

During training, BERT takes in a sequence of tokens as input and produces a sequence of hidden representations, one for each input token. These hidden representations are then used for downstream tasks such as classification, question-answering, and text generation. Since the bidirectional nature of BERT comes from its use of masked self-attention, where each token can attend to all the tokens in the input sequence, the training step does not require explicit bidirectional reading of the input sequence. However, the resulting hidden representations do capture bidirectional context information, making BERT a powerful language model for a wide range of natural language processing tasks. Additionally, BERT uses a technique called positional encoding to incorporate information about the position of each token in the input sequence, which helps the model distinguish between tokens with the same value but different positions.

Finally, in the context of molecule generation, conditioning can be used to guide a transformer model to generate molecules with desired properties, such as high potency or low toxicity. Conditioning refers to the process of incorporating additional information into a model to influence its output or behaviour. This additional information can be in the form of additional input data, constraints, or other relevant features that can guide the model to produce more desired outputs [12]. Conditioning is often used to improve the performance of a model or to make its output more interpretable or controllable, which can allow a chemical model to learn how to generate molecules that satisfy a given set of properties while still maintaining their structural and chemical diversity.

3 Research Questions

The proposed project aims to investigate several research questions related to the generation of molecules using deep learning models. Specifically, the following research questions will be addressed:

1. *How do the results of various models compare when trained on SMILES and SELFIES representations?*
   This question seeks to compare the performance of deep learning models that use the two molecular representations described earlier, namely SMILES and SELFIES. The objective is to validate the advantages and limitations of each representation and to identify the most suitable representation for the molecule generation task when using transformer-based language models.

2. *To what extent can language models based on the transformer architecture be used for the molecule generation task?*
   This research question aims to evaluate the effectiveness of transformer-based language models, such as BERT, for generating molecules. More specifically, the aim is to assess whether the trade-off between a higher computational cost and model performance is warranted, when compared to RNNs.

3. *Can a conditioning mechanism be used to constrain the generation process to a certain dataset identity?*
   Through this research question, this project seeks to explore the use of a conditioning mechanism to bias the molecule generation process towards a specific set of characteristics. Based on these properties, the objective is to investigate whether this approach can improve the diversity and quality of generated molecules.

4. *Can phasic update steps improve the quality of generated molecules?*
   By sampling the language model for a number of molecules every few steps, a set of metrics can be calculated (cf. Section 4.2) to verify how well a dataset distribution is learned. Using these results, this project aims to investigate the effectiveness of phasic update steps, which evaluate and update the model on additional criteria every few steps.

4 Methods

The process of *de novo* molecular design is a complex and challenging task, and its success heavily relies on the quality of the molecular representation and the machine learning models used. To ensure that the research questions are addressed, in this section, we describe methods used in this project to generate novel molecules with desired properties. The section begins by describing the datasets used in the experiments, followed by the metrics used to evaluate a given model. Finally, the models used in this project are described.

4.1 Datasets

To train the language models, three important datasets in the field of molecular design will be utilized, namely Kraken [8], OSCAR [7], and Tartarus [17]. Kraken is a discovery platform for monodentate organophosphorus (III) ligands that provides physicochemical descriptors based on representative conformer ensembles. OSCAR, on the other hand, is a repository of organocatalysts, building blocks, and combinatorially enriched structures that were experimentally derived. Lastly, Tartarus is a benchmarking platform for realistic and practical inverse molecular design, which provides tasks that rely on physical simulation of molecular systems to mimic real-life molecular design problems.

In total, the three datasets provide $\sim 975,000$ molecules in the SMILES format, which will be converted to the SELFIES representation. This process will be performed using the SELFIES package, an open-source Python library provided by [13] for molecular sequence representation. This library facilitates the encoding and decoding of molecules, which allows a language model to generate and optimize molecular structures based on any desired data distribution easily.

4.2 Metrics

There are several metrics that can be used to evaluate the performance of a language model in *de novo* molecular design. One of the most commonly used metrics is Quantitative Estimate of Drug-likeness (QED; 3), which provides a score that reflects the drug-like properties of a molecule based on its physicochemical and pharmacological properties. This metric is important because it enables the identification of molecules that have a higher probability of being successful drug candidates. Other important metrics are the Synthetic Accessibility Score (SA; 6), which estimates the ease of synthesis of a molecule based on its chemical structure. This metric is crucial because it enables the identification of molecules that are easy to synthesize, thus reducing the time and resources required to develop a promising compound.

These metrics, along with the Octanol-Water Partition Coefficient (Log P; 22), Bertz Complexity (BCT; 2), and Natural Product Likeness (NP; 5), can be calculated given a distribution of molecules. Other important features can be obtained using the MORFEOUS [11] library, such as pyramidalization and solvent accessible surface area. Using this information, an update step can be performed to improve the language model's performance. As it will be described shortly, this phasic behaviour can potentially be tuned to favour certain characteristics. Finally, the validation step will take into account the validity (only for SMILES), uniqueness, and novelty of the resulting molecules.

4.3 Models

Before training the language models on the SMILES and SELFIES datasets, a data pre-treatment step will be performed to ensure that the input sequences are in a suitable format. The SELFIES sequences will be pre-treated using the SELFIES package, which provides an efficient way to encode and decode molecular structures. The SMILES sequences will be pre-treated using RDKit [14], a cheminformatics library in Python that provides tools for handling and manipulating molecular structures.

After the pre-treatment step, both the BIMODAL and BERT [15] models will be trained on the SMILES and SELFIES data. To train the models, we will use the self-supervised learning approach, where the models are trained to predict missing tokens in a given sequence. The BIMODAL model will be trained to predict the next token in the SMILES and SELFIES sequences, while the BERT model will be trained to predict the masked tokens. Once the four models are trained for a given number of epochs, a comparison will be made between the SMILES and SELFIES-based models using the metrics presented in Section 4.2. Similarly, a comparison will be performed between the transformer-based language model and the RNN to identify which performs better.

Furthermore, as described earlier, this project explores the use of a conditional transformer model to retain the identity of the original dataset during the generation process. To achieve this, the CTRL model proposed by [12] will be trained on the three datasets described in Section 4.1. The control codes that will be provided during training represent the dataset identity, which are keywords that represent the content of each dataset (e.g., organophosphorus ligand, organocatalyst, organic photovoltaic). An attempt will be made to prepend these codes at the beginning of the molecule, or by injecting a representation of the dataset after the encoding process.

Finally, an ablation study will be performed, where updates based on a distribution of generated molecules are introduced every few steps. To this end, a cost function that takes into account the previously mentioned metrics will be defined to guide the optimization process. As described earlier, this phasic update step aims to improve the diversity and uniqueness of generated molecules. Furthermore, we investigate the effect of varying the frequency of the phasic update step on the resulting metrics, and compare these results against a model trained without phasic updates.

5 Scientific Relevance for Artificial Intelligence

The scientific relevance of this project lies in the application of deep learning models to molecular design, specifically the use of transformer-based language models to generate novel molecules with desired properties. The project aims to explore the use of different molecular representations and conditioning mechanisms to improve the diversity and quality of generated molecules. In the context of drug discovery, this project's relevance lies in the ability to generate novel molecules with desired properties, which can accelerate, for instance, drug development, catalyst design, or material design.

The project's contributions can advance state-of-the-art techniques for *de novo* molecular design, which can help identify molecules with desirable properties, reducing the need for costly and time-consuming experimental screening. Furthermore, evaluating the ideas proposed in Section 4 can provide insight into ways of improving chemical language models. This step can open up new opportunities for applications in chemistry and chemical engineering.

6 Planning

As it can be observed in Figure 1, the project is divided in 7 work periods (WP). In the first period (WP1), the literature behind the project is reviewed, and the project proposal is written. In WP2, the proposed molecular datasets are obtained, and their content is converted to the SMILES and/or SELFIES representations. In WP3, the BIMODAL model of [9] is trained on the three datasets obtained in WP1, for both SMILES and SELFIES. Simultaneously, WP4 will focus on training the BERT model on the obtained datasets, and attempts will be made to implement the phasic update steps. During WP5, the conditional transformer will be trained and evaluated, and final results will be compiled. Finally, WP6 represents a buffer, after which the final thesis will be written in WP7.

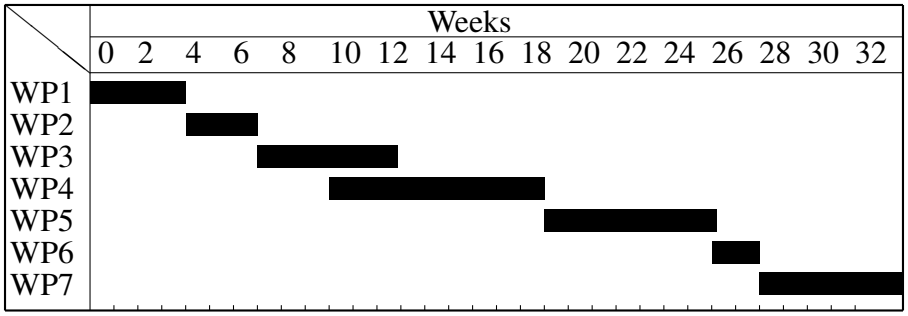

Figure 1: Duration of Work Periods

7 Resources and Support

In addition to the supervisors that provide expertise in machine learning and molecular science throughout the project, there are several other resources available to ensure its success. The University of Groningen provides access to computational resources such as Hábrók, the compute cluster provided by the Center for Information Technology. This high-performance computing cluster provides the computational power needed to run large-scale simulations and train complex machine learning models. Furthermore, the project will leverage open-source software libraries and frameworks, such as PyTorch and RDKit, which have been widely adopted by the scientific community for molecular modelling and machine learning tasks. These libraries offer a range of tools and functions for data pre-processing, model training, and evaluation, allowing for efficient and scalable implementation of the proposed methods. Finally, to ensure reproducibility and facilitate dissemination, the project will be developed following best practices in open science, including version control using Git and publication of code and data on public repositories such as GitHub.

References

[1] J. Arús-Pous, S. V. Johansson, O. Prykhodko, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen, and O. Engkvist. Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics*, 11(1):71, Nov. 2019.

[2] S. H. Bertz. The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601, June 1981.

[3] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, Jan. 2012.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] P. Ertl, S. Roggo, and A. Schuffenhauer. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. *Journal of Chemical Information and Modeling*, 48(1):68–74, Jan. 2008.

[6] P. Ertl and A. Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1:8, 2009.

[7] S. Gallarati, P. van Gerwen, R. Laplaza, S. Vela, A. Fabrizio, and C. Corminboeuf. OSCAR: An extensive repository of chemically and functionally diverse organocatalysts. *Chemical Science*, 13(46):13782–13794, Nov. 2022.

[8] T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, and A. Aspuru-Guzik. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *Journal of the American Chemical Society*, 144(3):1205–1217, Jan. 2022.

[9] F. Grisoni, M. Moret, R. Lingwood, and G. Schneider. Bidirectional Molecule Generation with Recurrent Neural Networks. *Journal of Chemical Information and Modeling*, 60(3):1175–1183, Mar. 2020.

[10] W. Jin, R. Barzilay, and T. Jaakkola. Junction Tree Variational Autoencoder for Molecular Graph Generation, Mar. 2019.

[11] K. Jorner and L. Turcani. kjelljorner/morfeus: v0.7.2, Aug. 2022.

[12] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher. CTRL: A Conditional Transformer Language Model for Controllable Generation, Sept. 2019.

[13] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, Oct. 2020.

[14] G. Landrum, P. Tosco, B. Kelley, Ric, sriniker, gedeck, D. Cosgrove, R. Vianello, NadineSchneider, E. Kawashima, D. N, A. Dalke, G. Jones, B. Cole, M. Swain, S. Turk, AlexanderSavelyev, A. Vaucher, M. Wójcikowski, I. Take, D. Probst, V. F. Scalfani, K. Ujihara, g. godin, A. Pahl,

F. Berenger, JLVarjo, jasondbiggs, strets123, and JP. Rdkit/rdkit: 2022_09_5 (Q3 2022) Release. Zenodo, Feb. 2023.

[15] O. Mahmood, E. Mansimov, R. Bonneau, and K. Cho. Masked graph modeling for molecule generation. *Nature Communications*, 12(1):3156, May 2021.

[16] J. Meyers, B. Fabian, and N. Brown. De novo molecular design and generative models. *Drug Discovery Today*, 26(11):2707–2715, Nov. 2021.

[17] A. Nigam, R. Pollice, G. Tom, K. Jorner, L. A. Thiede, A. Kundaje, and A. Aspuru-Guzik. Tartarus: A Benchmarking Platform for Realistic And Practical Inverse Molecular Design, Sept. 2022.

[18] G. Schneider and U. Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649–663, Aug. 2005.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need, Dec. 2017.

[20] W. Wang, Y. Wang, H. Zhao, and S. Sciabola. A Transformer-based Generative Model for De Novo Molecular Design, Oct. 2022.

[21] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, Feb. 1988.

[22] S. A. Wildman and G. M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, Sept. 1999.