UNIVERSITY OF TURKU

Molecules as Words

Department of Mechanical and Materials Engineering

Bachelor's thesis

Author:
Rosa Kataja

15.5.2025

Turku

Bachelor's thesis

Subject: Materials Engineering
Author: Rosa Kataja
Title: Molecules as Words
Supervisor: MSc Matilda Sipilä
Number of pages: 29 pages
Date: 15.5.2025

FI:

Viime vuosien aikana koneoppimisesta on tullut tärkeä väline uusien molekyylien ja materiaalien generointiin sekä niiden ominaisuuksien ennustamiseen. Perinteisistä, paljon aikaa ja resursseja kuluttavista, kokeellisista tavoista poiketen koneoppimismallit ovat tehokkaita työkaluja molekyylien ominaisuuksien ennustamiseen ja generoimiseen. Mallien tehokkuus riippuu kuitenkin paljon siitä, ovatko käsiteltävät molekyylit koneyhteensopivassa muodossa esitettyjä.

Tässä tutkielmassa perehdytään erityisesti kahteen molekyylien tekstimuotoiseen esitystapaan, SMILESiin (Simplified Molecular Input Line Entry System) ja SELFIESiin (SELF-referencIng Embedded Strings), sekä siihen, miten niillä esitetään molekyylejä koneoppimismalleille sopivassa muodossa. SMILES on alan nykyinen standardi, mutta sillä on taipumus tuottaa virheellisiä molekyylejä, kun taas SELFIES tuottaa vain kemiallisesti päteviä molekyylejä ja on siksi lupaava vaihtoehto etenkin generatiivisiin tehtäviin.

Kahden esitystavan tarkastelun lisäksi tutkielmassa esitellään myös viimeaikaisia syväoppimismalleja, jotka hyödyntävät SMILESia ja SELFIESiä. Malleihin kuuluvat transformer-pohjaiset SMILES-BERT ja SELFormer sekä generatiiviset mallit NRC-VABS ja DeLa-DrugSelf. Mallit osoittavat kemiallisten kielten käyttökelpoisuuden yhdistettynä nykyaikaisiin laskennallisiin menetelmiin.

Avainsanat: molekyylien tekstimuotoiset esitystavat, SMILES, SELFIES, neuroverkot, syväoppiminen, koneoppiminen, lääkeaineiden suunnittelu, molekyylien ominaisuuksien ennustaminen, molekyyligenerointi, generatiiviset mallit, luonnollisen kielen prosessointi (NLP)

EN:

In recent years, machine learning has become a key tool for advancing the discovery and optimisation of molecules and materials. Unlike traditional time-consuming experimental methods, machine learning models effectively predict molecular properties and generate novel molecules. Their effectiveness, however, depends on whether molecules are represented in a machine-readable format.

This thesis examines the use of molecular string representations to encode molecules for machine learning applications, focusing on SMILES (Simplified Molecular Input Line Entry System) and SELFIES (SELF-referencIng Embedded Strings). While SMILES is the current standard of the field, SELFIES is a promising candidate especially for generative tasks. The main issue with SMILES is the propensity to create chemically or syntactically invalid molecules. SELFIES was developed to address these issues and as a result it is a 100 % robust representation.

In addition to evaluating the two different string representations, this thesis also reviews recent deep learning models that utilise the two representations. These include transformer-based models SMILES-BERT and SELFormer, and generative models NRC-VABS and DeLa-DrugSelf. These models highlight the value of combining chemical languages with modern computational approaches.

Key words: molecular string representations, SMILES, SELFIES, neural networks, deep learning, machine learning, drug design, molecular property prediction, molecule generation, generative models, natural language processing (NLP)

Table of contents

1 Introduction

Developing new materials and molecules through experimental methods is often a time-consuming and resource-intensive process. Laboratory work includes substantial trial and error and requires costly reagents and specialised equipment, which slows down both the research process and progress in fields such as drug discovery and materials engineering. Traditional laboratory-based research methods remain essential, but modern technology offers alternative and complementary methods to them. Increasing availability of computational power and data-driven approaches such as machine learning accelerates the discovery and optimisation of materials.

Computational methods can be applied to a wide range of tasks, including property prediction, the identification of promising molecules and the generation of entirely new molecules. However, for machine learning models to effectively perform in their materials chemistry-related tasks, the molecules must be represented in a way that is both machine-readable and chemically correct.

Throughout history scientists have developed numerous representations for molecules to store structural information and facilitate communication within the field. These representations include visual formats like 2D drawings and 3D models, as well as various text-based notations. While these formats work well for human users, most of them aren’t suitable for machine learning models. Many were not designed with machine learning in mind and might lack the consistency or structure needed for computational processing.

In order to utilise the full capabilities that machine learning offers, molecular representations must be both chemically accurate and computationally compatible. These needs have led to the development and utilisation of molecular string representations such as SMILES and SELFIES, which encode molecules as character sequences in a way that is interpretable for machines. These representations make applying natural language processing (NLP) techniques to chemical data possible, while also expanding the possibilities of modern automated analysis and molecule generation.

2 Molecular string representations

The usage of machine learning (ML) and artificial intelligence (AI) in materials sciences and computational chemistry is growing rapidly, increasing the need for machine readable representations of molecules. String representations provide one approach to representing molecules in the age of AI and ML, alongside other representations such as molecular graphs and adjacency matrices. These string-based formats are based on the molecular graph theory, where molecules are considered as chains of atoms written as letter sequences in a string. Compared to other molecular representations such as graph representations, string representations are also relatively easy for humans to read and learn. [1]

Creating representations for small and simple molecules is relatively straightforward. However, as molecular structures become more complex, these representations must also become more expressive. This complexity introduces challenges in balancing syntactic validity, chemical accuracy and machine interpretability. The growing number of AI-based applications using string representations, such as molecular property prediction and generation of novel molecules, further increases the demand for expressive and robust molecular representations. [1, 2]

To meet this demand, researchers have developed various string representations over the years. In 1949 IUPAC issued a call for proposals because there was a growing need for an international notation system that used only ASCII letters and would be simple enough for typewriters and printing presses. Out of the multiple proposals, Dyson's ciphering was chosen as the standard but others were also used, such as the Wiswesser Line Notation (WLN) which became more popular in the 1960s. IUPAC's proposal laid the foundation and direction for modern string representations.[1] In 1988, David Weininger, a researcher interested in chemical databases, published SMILES (Simplified Molecular Input Line Entry System) which later became the de-facto standard for computational chemistry. However, machine learning and the use of AI to generate new molecules has introduced a problem that SMILES creates a large quantity of invalid molecules. DeepSMILES was introduced to address syntactical issues, but it fails to resolve chemical invalidity, such as unusual valence structures. In 2020, about 30 years after the introduction of SMILES, SELFIES (SELF-referencIng Embedded Strings) was introduced. SELFIES is a 100 % robust string representation model, meaning that all SELFIES strings are valid, and every molecule can be represented. [3]

2.1 SMILES

SMILES (Simplified Molecular Input Line Entry System) was introduced in 1988 by David Weininger. It was designed to be a truly computer interactive chemical notation language that remains user-friendly and easy to learn. By the 1988, computer technology had advanced to a point where minimising the length of molecular notations was no longer necessary. As a result, the use of complex rules and restricted symbols to shorten notations, a common feature in earlier representations, could be abandoned, allowing for improved readability. Consequently, SMILES and its rules are easy to interpret for both humans and machines. [4]

In SMILES, each atom is represented by its atomic symbol where the first letter written in upper case and the second in lower case. Aromatic atoms are written in lower case to differentiate them from non-aromatic ones. Hydrogen atoms are typically omitted, as they can be inferred based on standard valency rules. This means that hydrogens are excluded unless their count is unusual or required by other rules. Negative and positive charges are denoted within square brackets alongside the atom, and in these cases, attached hydrogens must also be specified. Single and aromatic bonds are often omitted but can be represented using the symbol "-" for single bonds and ":" for aromatic bonds. Double and triple bonds are represented by "=" and "#". Figure 1 illustrates a few example molecules along with their corresponding SMILES notations, demonstrating the application of the language's rules. [4]

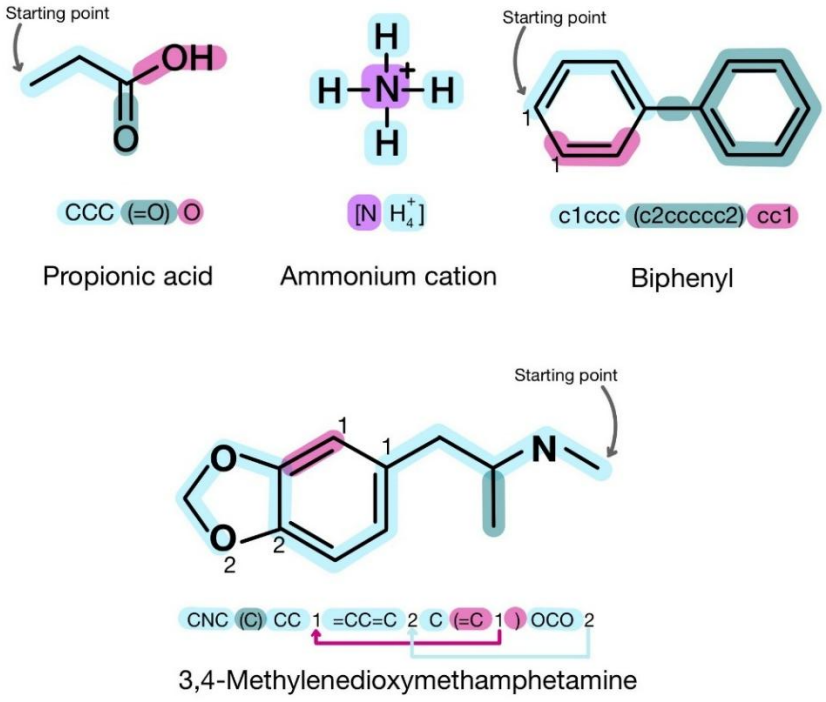


Figure 1. Example molecules and their SMILES representations.

The downside of SMILES is the ability to create invalid molecules making it hard to utilise for many applications. Generated SMILES may be invalid because of syntax errors such as conflicting binding symbols, most often meaning a missing or an extra branch or ring closure, or due to violating common chemical rules. When SMILES strings are tokenized, the possible large distances between branch and ring openings and closings can be a cause struggle to many neural networks. [5]

2.2 SELFIES

To address the issues and limitations of earlier molecular string representations, especially SMILES, SELFIES (SELF-referencIng Embedded Strings) was introduced in 2020. Before the issues, such as syntactical invalidity, were addressed by adapting machine learning models to handle errors. However, these approaches failed to resolve the main problem. SELFIES provides a solution by offering a 100% robust molecular string representation, where every possible SELFIES string corresponds to a valid molecule. It can be directly used by machine learning models without requiring any modifications to the language.

SELFIES grammar has derivation rules which produce each symbol. These symbols are written inside brackets, and the start of structures like rings and branches are defined using symbols such as “[Ring1]” and “[Branch1]”. The number following a ring symbol defines the size of the ring and is overloaded. Similarly to SMILES, SELFIES uses atomic symbols to represent atoms and the same symbols for different types of bonds. A comparison between the SMILES and SELFIES representations of 3,4-methylenedioxymethamphetamine (MDMA) is illustrated in Figure 2, demonstrating the rules of SELFIES. [1]

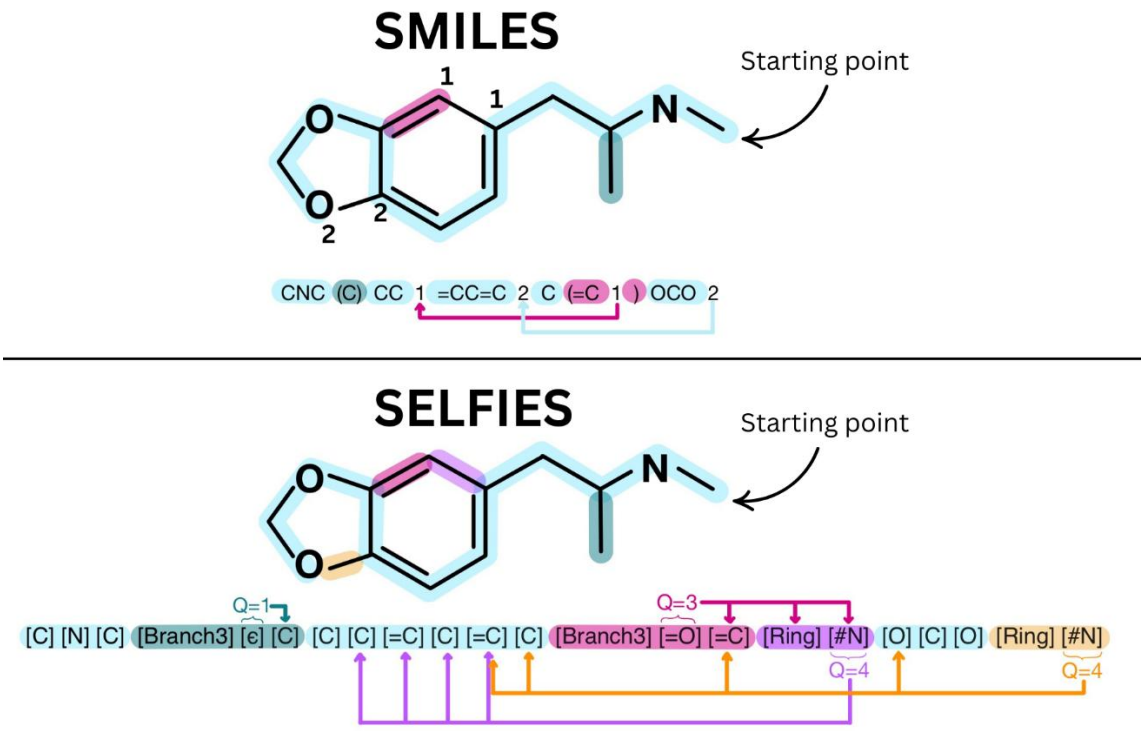


Figure 2. SMILES and SELFIES representations of 3,4-methylenedioxymethamphetamine.

One of the most significant advantages of SELFIES is its robustness: even randomly generated strings or mutated strings always correspond to valid molecules. In contrast, SMILES strings are prone to becoming invalid with even small mutations. As illustrated in Figure 3, when random mutations are applied to the MDMA molecule, the resulting SMILES string are invalid, whereas SELFIES strings remain valid.

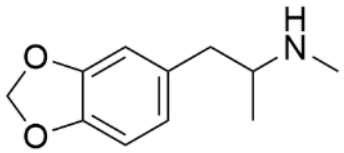
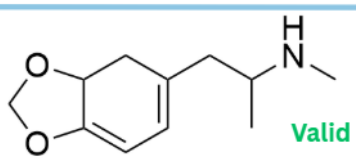
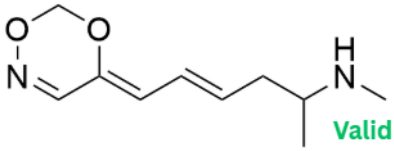
SMILES	SELFIES
Original Molecule **CNC(C)CC1=CC=C2C(=C1)=OCO2**	
Single Mutation **CNC(C)CC1=CC=CNC(=C1)=CO2** **Invalid**	 **Valid**
Double Mutation **CNC(C)OC1=CC=C2C(=C1COCO2** **Invalid**	 **Valid**

Figure 3. Applying random mutations to MDMA produces invalid SMILES but completely valid SELFIES molecules. Mutations are shown in red.

Although SELFIES is generally considered less readable for humans than SMILES, it is still easy to implement after getting familiar with the language, especially given its similarity to former molecular string representations. [3] Currently, SELFIES can represent common organic molecules including chirality and stereochemistry. SELFIES is being developed to be even more versatile which would mean being able to represent macromolecules, crystal structures and complicated bonds. [1]

3 Neural Networks

Neural networks are computational models inspired by the structure and function of the human nervous system [6]. Like biological neural networks, Artificial Neural Networks (ANNs) consist of hundreds of interconnected neurons, or nodes, which process and transfer input data through weighed connections [7]. In addition to receiving data, the nodes can also store information which affects the weighed output [6]. In recent years, neural networks have been increasingly applied in various scientific fields including chemistry and materials science. The ability to model complex relationships in data makes neural networks particularly useful for property prediction.

ANNs utilise two primary training approaches: supervised and unsupervised learning. In supervised learning, the model is provided with input data along with corresponding outputs, and its parameters are adjusted based on the learning results. In unsupervised learning, only input data is given, and the model must identify patterns or trends on its own. The weighed connections between nodes are adjusted based on the training results to optimise model performance. [6]

Neural networks consist of different layers of nodes, typically divided into three categories based on their tasks: input, hidden and output layers. Figure 4 illustrates a basic neural network: data is received at the input layer, relationships between input and output are determined in hidden layers, and the resulting output is produced at the output layer. Depending on the neural network, there may be one or multiple hidden layers. Traditional ANNs process data unidirectionally meaning that data flows through layers without feedback loops. [6, 7]

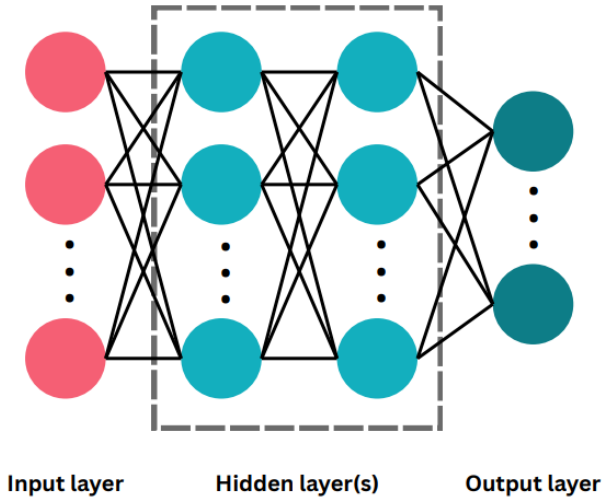


Figure 4. Example of a neural network, including input and output layers and two hidden layers.

Recurrent Neural Networks (RNNs) differ from traditional feedforward ANNs by incorporating feedback connections in their hidden layers. This architecture allows RNNs to store, remember and process previous data for extended periods. Due to their structure, RNNs are particularly effective at modelling sequential data and acknowledging sequences. Training RNNs typically follows a supervised approach, requiring input-target pairs during training. [8, 9]

Variational Autoencoders (VAEs) are commonly used generative models that represent data through a latent space and are trained in an unsupervised manner. Latent variables capture lower dimensional data than the observed data, allowing the model to encode the data in a way that enable new data to be generated from the latent variables. VAEs consist of two main components: an encoder, which compresses the data into a latent space, and a decoder, which reconstructs data samples from the latent space. [10]

BERT (Bidirectional Encoder Representations from Transformers) is one of the most advanced embedding models, introduced by Google in 2018. BERT's distinguishing feature is its ability to consider contextual relationships, meaning it processes words in relation to all other words in a sentence, leading to improved classification and accuracy. The model is pre-trained using two tasks: masked language modelling (MLM) and next sentence prediction (NSP). In masked language modelling 15 % of words in a sentence are masked and the model is asked to predict the masked words. In NSP the model is given two sentences, and it must figure out whether the sentences are consecutive. [11, 12]

A common challenge in machine learning is overfitting, which occurs especially in supervised learning tasks. Machine learning models are trained on certain datasets in order to make predictions on new data, but if a model learns the training data too well it fails to find general predictive rules. Learning the data too well would mean not only capturing the patterns within data but also noise and random fluctuations in the data. Overfitting is often caused by applying an objective function that prioritises minimising errors in training data rather than generalisation, which leads to memorising peculiarities in training data. There are several methods to mitigate overfitting such as regularisation functions and cross-validation, which modify the objective functions to balance fitting the training data and generalising new data. Additionally, simple algorithms such as greedy search and gradient descent can help reduce overfitting. [13]

4 Applications

4.1 SMILES-BERT

SMILES-BERT, introduced in 2019, is a semi-unsupervised deep learning model designed for predicting molecular properties of materials. It leverages BERT architecture and uses SMILES strings to represent molecules. This approach enables the model to learn molecular properties from large-scale unlabelled data before fine-tuning on smaller labelled datasets. [14]

Because SMILES symbols aren't consecutive the second pre-training task included in BERT, NSP, isn't applicable in SMILES-BERT. With masked language modelling the selected training datasets for SMILES-BERT grow larger which helps with overfitting. By excluding NSP, the base of BERT in this model only has 6 transformer encoder layers instead of the usual 12 layers and requires less computation and memory. A larger SMILES-BERT model containing 12 layers was also created and compared with the 6 layer one (see Table 1). Both models significantly outperformed other state-of-the-art methods in accuracy but offered no notable improvements in performance; therefore, other structures were not trained. As shown in Table 1, the two versions, SMILES-BERT and SMILES-BERT (large), achieved nearly identical accuracy results. Due to the larger model's significantly higher training time without better results, the smaller version of SMILES-BERT was chosen for further evaluation. [14]

Table 1. Structural differences and performance comparison of the two SMILES-BERT models.

	layers	att-heads	accuracy
SMILES-BERT	6	4	**0,9154**
SMILES-BERT (large)	12	12	0,9147

The dataset of SMILES molecules used for pre-training is from a free database called ZINC that includes 35 million compounds. Pre-training is conducted without additional labels to the SMILES labels. Setting the first 4000 training steps as warm-up was found to be important for the model to avoid not converging even after a long time. The model was able to correctly recover 82,85% of the masked SMILES during pre-training. [14]

Evaluation of the model was conducted using three datasets, PM2, PCBA and LogP, differing in size and molecular properties. However, detailed information of the PM2 dataset and its molecular content was not available. Properties of the datasets are summarised in Table 2. SMILES-BERT was compared to four other advanced methods, Circular Fingerprint, Neural Fingerprint, Seq2seq Fingerprint and Seq3seq Fingerprint. SMILES-BERT had the best performance on all three datasets compared to the four other methods. Due to the lack of information of the PM2 dataset, the comparison here only focuses on the LogP and PCBA datasets. The most significant result came from the LogP dataset, which contained over 10 000 SMILES and water-octanol partition coefficient (LogP) pairs, where SMILES-BERT achieved a prediction accuracy of 91,54% which was about 2% better than the second-best method. Comparison of the results from the LogP dataset is in Figure 5. The results of the second largest dataset, PCBA-686978, are represented in Figure 6, revealing that SMILES-BERT performed about 3% better than the second-best method, achieving an accuracy of 87,84%. [14]

Table 2. Datasets used for evaluation of SMILES-BERT, along with their sizes, accessibilities and descriptions. [15]

	amount of samples	publicity	description
LogP	>10 850	nonpublic	water-octanol partition coefficients
PM2	323 242	nonpublic	-
PCBA-686978	302 175	public	bio-activity of small molecules

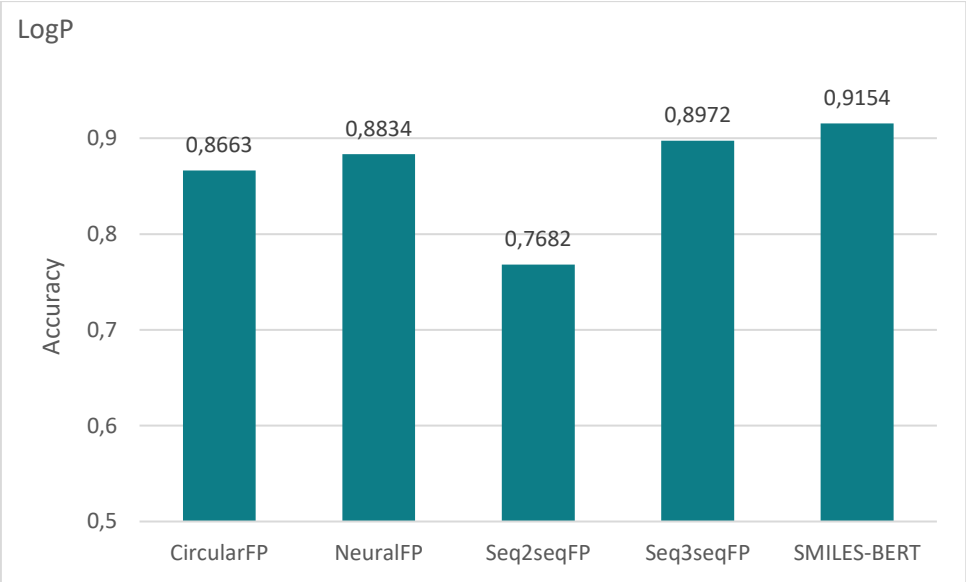


Figure 5. Comparison of five different models and their accuracy results on the LogP dataset.

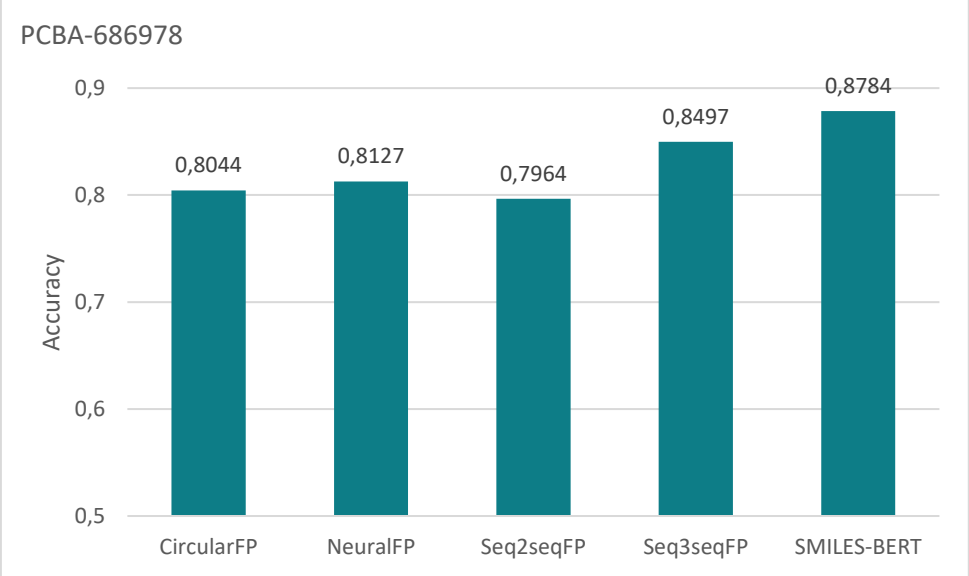


Figure 6. Comparison of five different models and their accuracy results on the PCBA-686978 dataset.

SMILES-BERT demonstrates strong performance in molecular property prediction using a semi-supervised learning approach based on BERT architecture. By training on large-scale data using MLM and fine-tuning on labelled datasets, the model achieves superior results compared to other state-of-the-art models. Results from evaluation datasets, PM2, PCBA and LogP, confirm that SMILES-BERT has strong predictive capability and generalisation potential. Future work could include adding Quantitative Estimate of Druglikeness (QED) as an additional task in pre-training to increase the classification capability.

4.2 SELFormer

SELFormer is a chemical language model built on transformer architecture that uses SELFIES as input. It's designed for molecular property prediction and is trained on large-scale data before fine-tuning for various specific prediction tasks. When SELFormer was published in 2023, it outperformed competing methods on several property prediction tasks while delivering comparable results in others. This highlights its effectiveness in using SELFIES representations for chemical modeling. [16]

Because most molecular datasets still rely on SMILES notations, an important preprocessing step for the training data of SELFormer involves converting these representations into SELFIES. Over 2 million drug-like bioactive compounds from ChEMBL dataset were chosen for training and SELFIES conversion. SELFormer utilises RoBERTa in pre-training which is based on BERT architecture. RoBERTa excludes the NSP task in pre-training and only uses MLM. The steps included in pre-training are shown in Figure 7. Instead of training thousands of models, only a hundred models were trained to save resources. Out of the 100 pre-trained models the best two were selected for further pre-training and fine-tuning. The two chosen models are referred to as SELFormer Lite (less trainable parameters) and SELFormer. Both models are pre-trained on the full ChEMBL dataset. [16]

Fine-tuning was conducted using multiple datasets from MoleculeNet to train the model for different property prediction tasks. The datasets used for classification tasks were BBBP (Blood-Brain Barrier Penetration), SIDER (The Side Effect Resource), Tox21 (Toxicology in the 21st century), HIV (ability to inhibit HIV replication) and BACE (binding properties against the human beta-secretase 1 protein). For regression tasks the model used FreeSolv (Free Solvation Database), ESOL (aqueous solubility) and PDBbind (Lipophilicity and the binding affinity prediction). As Figure 7 shows, in fine-tuning the embedding of the pre-trained model is taken as input and mapped out by a classification/regression head to the corresponding output. For classification-based tasks results are given as ROC (area under receiver operating characteristic curve) and for regression-based tasks results are given as RMSE (root mean squared error). [16]

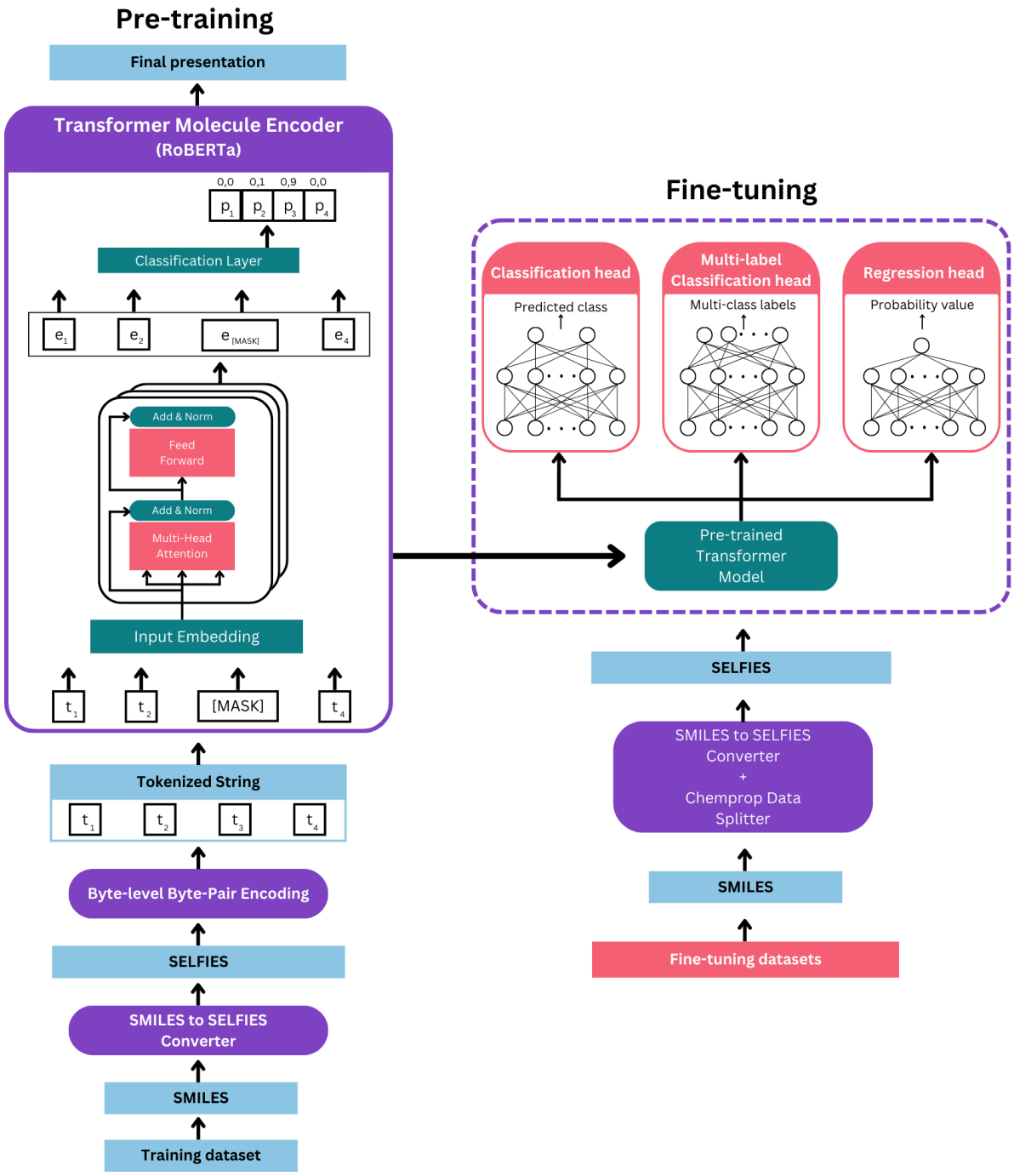


Figure 7. Overview of the SELFormer architecture. In pre-training the SMILES strings are first converted into SELFIES and then tokenized using byte-level byte-pair encoding. These tokens are used as input for a RoBERTa-based transformer encoder using masked language modeling. During fine-tuning, the pre-trained model is adapted for various tasks, including multi-label classification and regression. Converting SMILES strings into SELFIES representations is also needed in fine-tuning.

When compared to ten other graph- and text-based models on nine tasks, SELFormer outperformed others only on two tasks, ESOL (aqueous solubility), a regression task, and

SIDER (Side Effect Resource), a multi-label classification task. Both datasets were relatively small, each including under 1500 compounds. As Figure 8 illustrates, other models performed about 10 % worse on the SIDER task, while Figure 9 shows that their performance was approximately 15 % worse on the ESOL task. [16]

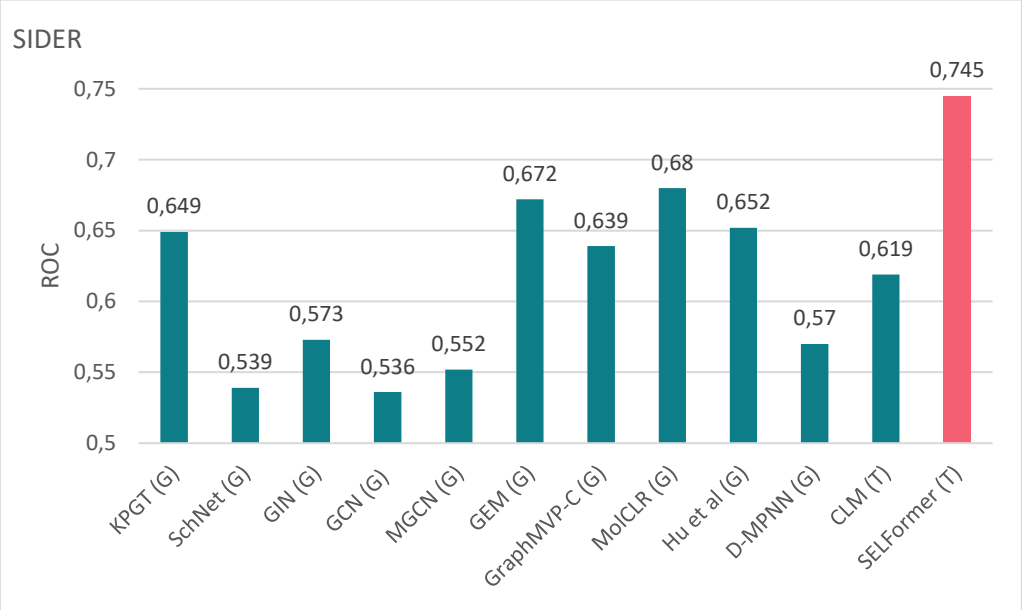


Figure 8. Performance of different text- and graph-based models on the SIDER task. Results are given as ROC scores. Models labelled with (G) are graph-based, while the ones with (T) are text-based.

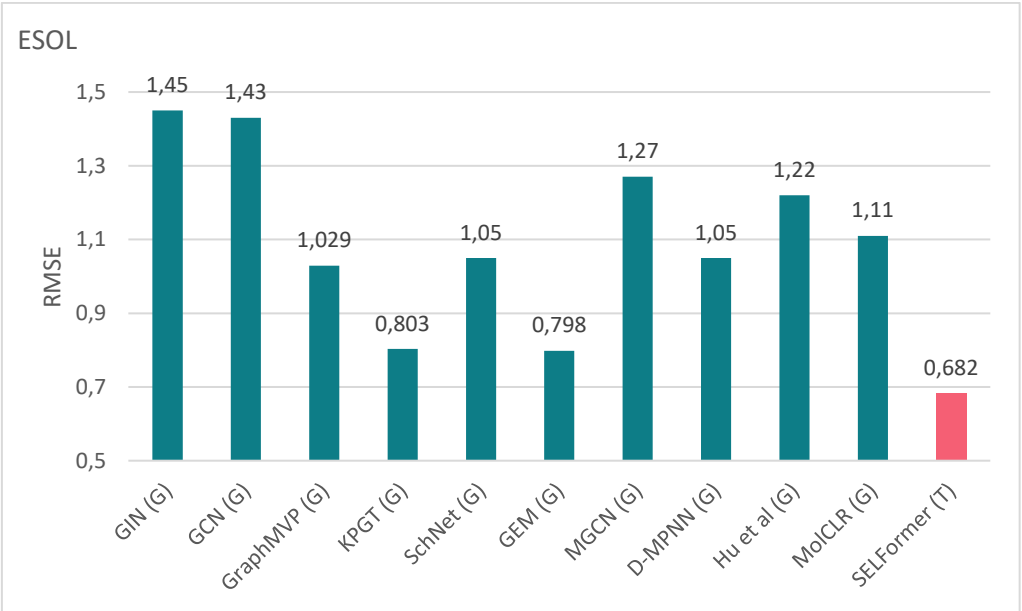


Figure 9. Performance of different text- and graph-based models on the ESOL task. Results are given as RMSE scores. Models labelled with (G) are graph-based, while the ones with (T) are text-based.

SELFormer achieved significantly better results compared to most graph-based models in molecular property prediction tasks. ChemBERTa-2 provides a notable point of comparison, as it uses the same BERT architecture as SELFormer but instead of SELFIES uses SMILES notations. When comparing the two models, SELFormer outperformed on all tasks which suggests that SELFIES may be more suitable for property prediction models. The fact that SELFIES is able to represent a wider range of molecules than SMILES could be the cause of this performance difference. [16]

4.3 NRC-VABS

Most LSTM (long short-term memory) based VAE deep learning models designed for the generation of drug molecules struggle with posterior collapse [17]. Posterior collapse occurs when the learned latent variables disregard the meaningful information from input data. This typically occurs when the decoder is too powerful and is able to generate valid outputs without relying on the latent space. As a result, the generated molecules are less diverse and fail to depend on the input data. [18] To overcome this NRC-VABS (Normalized Reparameterized conditional VAE with applied Beam Search in latent space), published in 2024, introduces a novel approach that improves molecular generation by tailoring new molecule samples to specific property constraints. The model takes SMILES representations as input but addresses the issue of invalidity in representations by converting them into normalised SMILES formats, named H_XSMILES. [17]

Modification of the SMILES representation into a normalized version results in a less complex string and for that reason easier to use for the DL model. H_XSMILES modifies the pair representations in SMILES, including branches and rings. Instead of placing a set of parentheses at both ends of a branch, the normalized model uses a closing bracket followed by a number to indicate the presence of a ring and the length of it. Ring structures are represented by adding a "^"-symbol after the last ring molecule, followed by a number denoting the size of the ring and another number after a "_"-symbol specifying the bond between the first and last atoms in the ring. Examples of the differences between SMILES and H_XSMILES molecules are represented in Figure 10. [17]

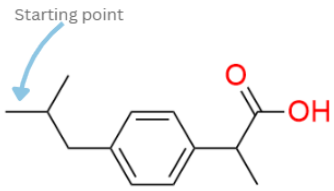
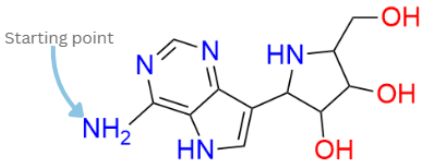
	Ibuprofen	Galidesivir
		
SMILES	CC(C)Cc1ccc(cc1)C(C)C(O)=O	Nc1ncnc2c(C3NC(CO)C(O)C3O)c[nH]c12
H_xSMILES	CCC)CC=CC=CC=C^6_1)2CC)1CO)1=O	NCN=CN=CCCNCCO)2CO)1C^5_1O)6=C[NH]C^9_2^5_1

Figure 10. Two common drug molecules with their SMILES and H_xSMILES representations.

LSTM based models also provide good properties for the generation of SMILES molecules. The models can store long-term dependencies in sequential data and take inputs of diverse lengths. Combining LSTMs with the organization of VAE allows models to convert chemical property encodings into a lower-dimensional latent space that can be used for the generation of new molecules. [17]

Figure 11 illustrates the overall architecture of NRC-VABS, which incorporates two subnetworks, the encoder and decoder, both including three layers with each layer having 512 cells. Instead of using a common method called greedy search (GS) to explore the latent space, NRC-VABS uses beam search (BS). Compared to greedy search BS explores multiple paths, leading to more diverse generated samples. The model is trained with 250K and MOSES datasets both including drug-like molecules. After BS the model uses a configurable diversity parameter, as its absence would result in generated molecules being similar to the target molecules. [17]

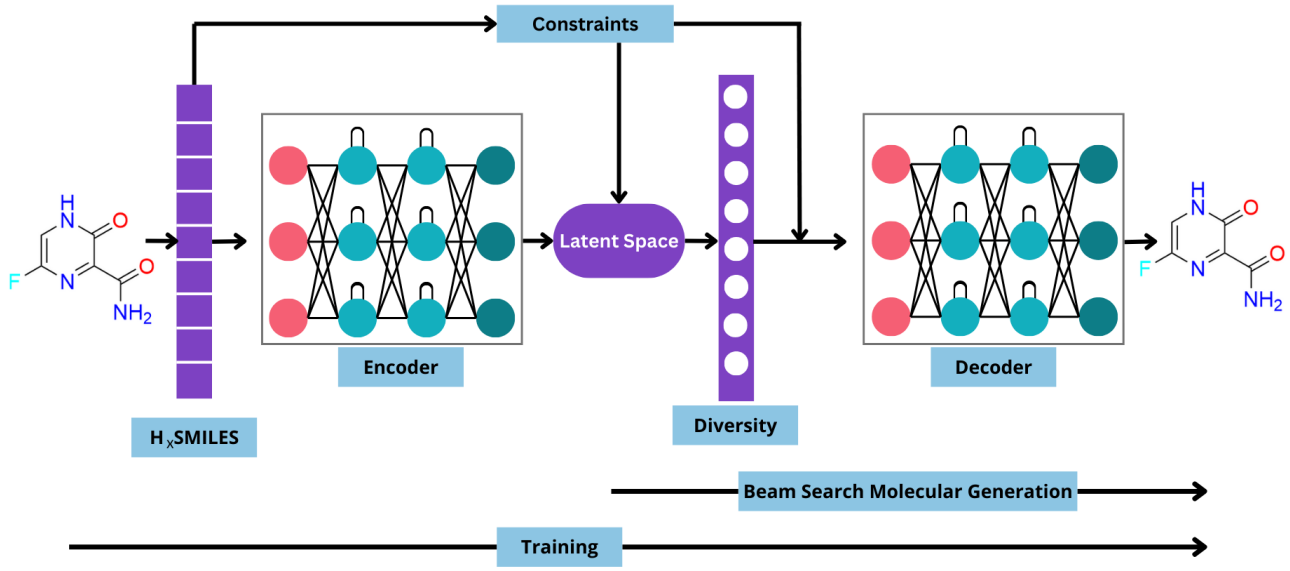


Figure 11. Architecture of NRC-VABS, including modification of SMILES into normalized versions, two subnetworks and beam search applied for the exploration of latent space.

The results of the 250K dataset were examined comparing four versions of NRC-VABS with differing diversifications to four other state-of-the-art models all utilising VAE and greedy search. Methods were compared on their validity, novelty, accuracy and uniqueness. The comparison metrics used were developed by the authors themselves and for example reconstruction accuracy was calculated as the ratio of correctly predicted characters to the total number of characters in the string (Formula 1). [17]

$$Reconstruction\ accuracy = \frac{\#\ matched\ characters}{\#\ characters\ in\ the\ molecules} \quad (1)$$

All NRC-VABS models outperformed the others on accuracy and almost all on validity when compared using the 250K dataset. Table 3 lists the accuracy results of different models and their selection methods. Novelty and uniqueness were only compared to one other model but within the NRC-VABS models uniqueness increased with diversification and novelty on the other hand decreased. Overall, all NRC-VABS models achieved great results with the average of the results being about 90 %. [17]

Table 3. Reconstruction accuracy results of NRC-VABS models and other state-of-the-art models and their selection methods.

	Selection	Accuracy
VVAE	GS	0,19
CVAE	GS	0,44
GVAE	GS	0,53
SD-VAE	GS	0,76
NRC-VAE$_1$	BS	0,89
NRC-VAE$_2$	BS	**0,97**
NRC-VAE$_3$	BS	0,96
NRC-VAE$_4$	BS	0,93

In addition to being a dataset, MOSES is also a platform that includes many state-of-the-art models and metrics that can be used to compare models. NRC-VABS was compared to nine other models, and it surpassed almost all of them in all categories. The models were compared on novelty, validity, uniqueness among other metrics. However, the performance differences between the models were minimal, typically ranging from under 1% to a few percentages. [17]

The main limitation of NRC-VABS is that it can only manage three properties and efforts to expand its capacity haven't yet been successful. Another limitation within NRC-VABS is that modifying one property often unintentionally affects the other because of the complex interdependencies among the molecular properties. [17] While NRC-VABS performed well on multiple tasks, these limitations are notable, as molecules fundamentally possess a wide range of properties, and many applications require the simultaneous consideration of multiple properties.

4.4 DeLA-DrugSelf

Drug discovery is a time-consuming process that involves modifying a known active molecule's structure slightly to enhance its activity against a target protein. Generative algorithms have improved the efficiency of the process, but because molecules are commonly represented as molecular graphs or SMILES, the issue of invalidity arises when using

SMILES. DeLA-DrugSelf (2024) is an improved version of DeLA-Drug (2022), and unlike the former version utilises SELFIES instead of SMILES. The model architecture is based on RNNs (Recurrent Neural Networks) and generates new molecules by substituting, deleting or inserting random characters in strings. [19]

Although SELFIES has many advantages, a frequently overlooked issue arises during the decoding: the algorithm automatically cuts the incorrect token sequences into valid SELFIES strings. This “collapse” problem can introduce a bias, lead to false positives in training and raise doubts about the accuracy of generative models. DeLA-DrugSelf addresses this problem by excluding collapsed strings from the generation, thereby improving the interpretability of the results. [19]

The model was trained using data that was adapted from the ChEMBL28 dataset. Several refinement steps were applied, including for example filtering out inorganic compounds, stereoisomerism and duplicates. Additionally, all compounds were converted from SMILES strings into SELFIES strings and tokenized. The RNN consists of two layers of LSTM units and is trained in two generative tasks: Sampling From Scratch (SFS) and Sampling With Mutations (SWM). In SFS, the model is trained to estimate the probability distribution of the following character based on the preceding characters as context. In contrast, SWM includes modifying a given SELFIES string by selecting a random number of tokens in random positions and applying one of three possible operations to them: insertion, substitution or deletion. The steps DeLa-DrugSelf follows to generate new molecules are illustrated in Figure 12. [19]

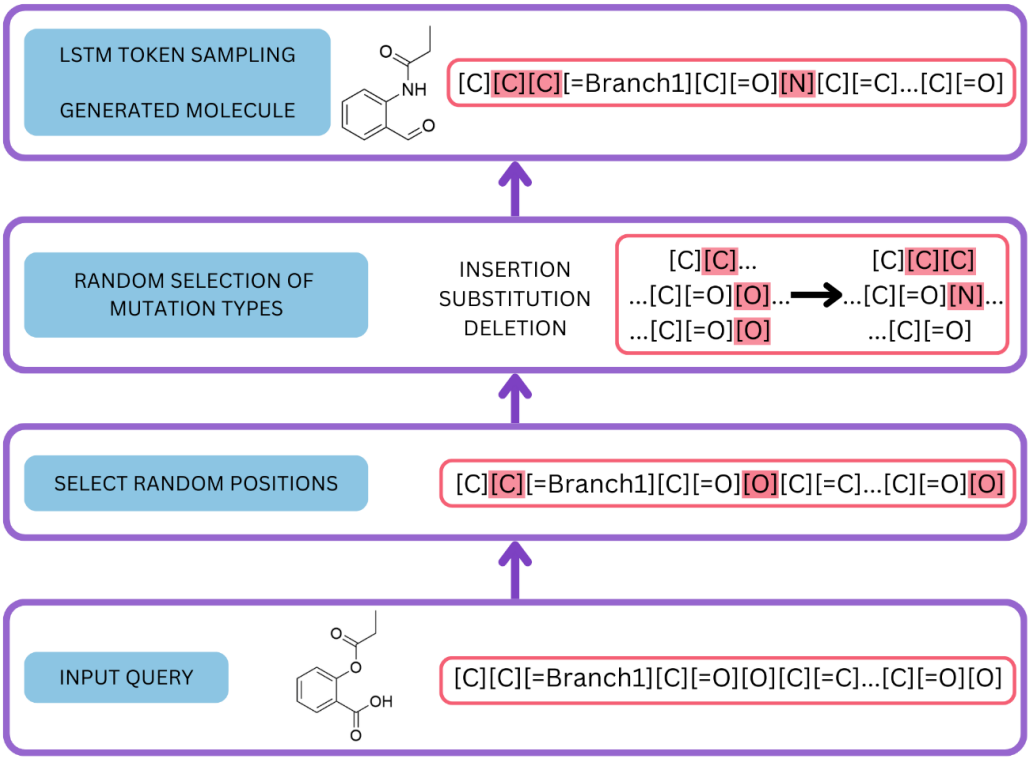


Figure 12. Main steps of the DeLa-DrugSelf architecture. Generation process includes insertion, substitution and deletion of parts of the string on randomly selected positions.

Results from the SFS approach were examined by generating 100 000 SELFIES strings under varying context lenghts. Metrics such as unicity, novelty, and the uncollapsed rate were used to evaluate quality. The uncollapsed rate exceeded 75% across most settings, while both unicity and novelty consistently surpassed over 99%. [19] These results significantly outperform the earlier DeLA-Drug model, which achieved unicity values below 89% using SMILES and the same SFS architecture. [20] This highlights the robustness of the SELFIES-based generation [19].

The SWM approach was evaluated by generating 100 molecules for each of 100 query compounds. A clear trend emerged: as the number of mutations increased from 1 to 5, the uncollapsed rate declined from 27% to 3%, reflecting the increased likelihood of syntactic disruption. Unicity results increased as the number of mutations increased, starting from just over 60%, increasing to 99%. Novelty results from all rates of mutation surpassed 99%. Considering only uncollapsed molecules, DeLA-DrugSelf outperforms DeLa-Drug in drug-likeness (QED). These results highlight the importance of managing collapsing and prove the

models potential in producing high-quality molecules closely related to reference molecules. [19]

In summary, DeLa-DrugSelf demonstrates a strong capacity for controlled and high-quality molecule generation. The main advantages of the model are the use of 100% robust SELFIES representations and the collapse-aware architecture. The combination of two generative tasks allows for broad exploration of chemical space, and results prove its potential as a versatile tool for drug design. Future improvements include adopting alternative architectures other than RNNS to further improve performance and scalability. [19]

5 Conclusions

Efficient representation of molecules has been a central research object for decades. However, as the range of applications for molecular representations has expanded, the requirements for chemical languages have also changed accordingly. While earlier string-based representations have proven to be compatible with machine learning, their limitations, particularly in validity and scalability remain significant.

The most widely used string representation remains SMILES, offering a compact and interpretable format to encode molecules, making it valuable for chemistry and materials science. However, in deep learning contexts, the high frequency of syntactically invalid strings and chemically invalid structures generated from SMILES often leads to significantly reduced model performance. These limitations motivated the development of alternative representations such as SELFIES, which ensures 100% syntactic and chemical validity. Despite being less human-readable, SELFIES enables more robust training, particularly in generative tasks.

Models such as SMILES-BERT and SELFormer demonstrate how transformer-based architectures can be fine-tuned to understand chemical syntax and predict molecular properties with high accuracy. Notably SELFormer's use of SELFIES instead of SMILES proved advantageous, resulting in comparable or improved performance across several tasks. These results suggest that SELFIES is a promising standard for future NLP-inspired chemical modeling.

In addition to property prediction, molecule generation is an increasingly relevant application of string-based representations. Models such as NRC-VABS and DeLa-DrugSelf aim to generate novel molecules that meet specific property criteria. The need in NRC-VABS to convert SMILES into normalised H_XSMILES before model input further highlights the limitations of SMILES for modern deep learning applications. In contrast, DeLa-DrugSelf leverages the robustness of SELFIES to avoid issues such as syntactic collapse and improve the quality of generated molecules.

In the future, promising directions of research include the development of hybrid models that combine the strengths of string- and graph-based representations. In addition, applying these models to more complex molecular structures – such as polymers, crystalline materials and inorganic compounds – could widen their applicability. Currently neither SELFIES nor

SMILES can fully accommodate such complex structures, indicating an important gap in current representation techniques.

Once the current limitations of string representations and machine learning models are resolved and the capabilities of these models expand toward unprecedented scales, multiple ethical considerations remain. It is a well-established that large deep learning models require substantial amounts of computational power and electricity, which raises concerns about the environmental impacts of these models. This highlights the need for political interest and collaboration between different scientific fields to balance predictive power with energy efficiency and sustainability.

Furthermore, the deployment of deep learning in critical fields such as drug discovery demands precise consideration of safety and accountability. Errors in generated molecules and predictions could have serious consequences, especially when applied to real-world medical and pharmaceutical decisions. While AI-generated drug molecules show promise, they will most likely require extensive testing by human experts to prove their safety and efficiency. This is important given the ongoing public uncertainty and suspicion about artificial intelligence and the fact that molecular interactions often involve complexities beyond the current capabilities of AI.

Although novel drug molecules designed by AI and other applications of molecular representations and deep learning could significantly improve the health of multiple humans, it’s crucial to consider the broader trade-offs. If the pursuit of these innovations leads to excessive resource consumption and environmental degradation, we must ask whether the long-term cost outweighs the benefit. As the scientist continue their ongoing work to develop AI-driven technologies, it is essential that this progress is guided by ethical responsibility in mind.

References

[1] M. Krenn *et al.*, SELFIES and the future of molecular string representations, *Patterns* 3 (2022), 100588, https://doi.org/10.1016/j.patter.2022.100588.

[2] M. E. Mswahili and Y.-S. Jeong, Transformer-based models for chemical SMILES representation: A comprehensive literature review, *Heliyon* 10 (2024) e39038. https://doi.org/10.1016/j.heliyon.2024.e39038.

[3] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation, *Mach. Learn. Sci. Technol.* 1 (2020) 045024. https://doi.org/10.1088/2632-2153/aba947.

[4] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* 28 (1988) 31–36. https://doi.org/10.1021/ci00057a005.

[5] K. Rajan, C. Steinbeck, and A. Zielesny, Performance of chemical structure string representations for chemical image recognition using transformers, *Digit. Discov.* 1 (2022) 84–90. https://doi.org/10.1039/D1DD00013F.

[6] A. D. Dongare, R. R. Kharde, and A. D. Kachare, Introduction to Artificial Neural Network, *Int. J. Eng. Innov. Technol.* 2 (2012) 189-194.

[7] D. J. Lvingstone, *Artificial Neural Networks: Methods and Applications*, Humana Press, Totowa NJ, USA, 2009, pp. 15-23. https://doi.org/10.1007/978-1-60327-101-1.

[8] L. R. Medsker and L. C. Jain, *Recurrent neural networks: Design and Applications*, CRC Press, Boca Raton FL, USA, 2001, pp. 13-18.

[9] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, Recent Advances in Recurrent Neural Networks, *arXiv* (2018), arXiv:1801.01078. https://doi.org/10.48550/arXiv.1801.01078.

[10] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, Dynamical Variational Autoencoders: A Comprehensive Review, *Found. Trends® Mach. Learn.* 15 (2021) 1–175. https://doi.org/10.1561/2200000089.

[11] S. Ravichandiran, *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT*. Packt Publishing, Birmingham, UK 2021, pp 53-90.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv* (2018), arXiv:1810.04805.

[13] T. Dietterich, Overfitting and undercomputing in machine learning, *ACM Comput. Surv.* 27 (1995), 326–327. https://doi.org/10.1145/212094.212114.

[14] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction, in *Proc. 10th ACM Int. Conf. on Bioinformatics, Comput. Biol. Health Inform.* (2019) 429–436. https://doi.org/10.1145/3307339.3342186.

[15] H. Li, X. Zhao, S. Li, F. Wan, D. Zhao, and J. Zeng, Improving molecular property prediction through a task similarity enhanced transfer learning strategy, *iScience* 25 (2022) 105231. https://doi.org/10.1016/j.isci.2022.105231.

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

[17] A. S. Bhadwal, K. Kumar, and N. Kumar, NRC-VABS: Normalized Reparameterized Conditional Variational Autoencoder with applied beam search in latent space for drug molecule design, *Expert Syst. Appl.* 240 (2024) 122396. https://doi.org/10.1016/j.eswa.2023.122396.

[18] Y. Wang, D. M. Blei, and J. P. Cunningham, Posterior Collapse and Latent Variable Non-identifiability, *arXiv*, (2019) arXiv:1904.07237.

[19] D. Alberga *et al.*, DeLA-DrugSelf: Empowering multi-objective de novo design through SELFIES molecular representation, *Comput. Biol. Med.* 175 (2024) 108486. https://doi.org/10.1016/j.compbiomed.2024.108486.

[20] T. M. Creanza *et al.*, DeLA-Drug: A Deep Learning Algorithm for Automated Design of Druglike Analogues, *J. Chem. Inf. Model.* 62 (2022) 1411–1424. https://doi.org/10.1021/acs.jcim.2c00205.