Exploring Tokenization Techniques, Denoisers, Metrics, and Training Strategies in AI-driven Molecular Design

Mahomed Aadil Ally

Department of Computer Science,
University of Cape Town,
Rondebosch, 7701,
South Africa
allmah002@myuct.ac.za

Abstract

This literature review explores the field of AI-driven molecular design, with particular attention paid to tokenization methods, denoisers, assessment metrics, and training approaches. Understanding these elements is essential for improving molecular generating processes, as there is growing interest in using AI models for drug discovery whilst still meeting obstacles in current developments. This study seeks to provide a thorough overview of current approaches and recommend opportunities for additional exploration by synthesizing ideas from existing research.

Introduction

The use of artificial intelligence (AI) in drug discovery has become a viable strategy in recent years to hasten the creation of innovative treatments. Because traditional drug discovery procedures are expensive and time-consuming, new approaches are needed to identify possible therapeutic candidates more quickly. An answer is provided by AI-driven molecular design, which makes use of computational models to forecast molecular attributes and produce candidate molecules with desired traits.

To train and use natural language processing (NLP) models, tokenization is a necessary preprocessing step for sequential data. Its implications on chemical applications, however, have not received enough consideration. Tokenization in the context of text production can have a major effect on prediction quality. It refers to procedures used in chemistry to disassemble linear molecule representations into their component parts. To make it easier to integrate complex chemical structures into machine learning models, linearization entails expressing them as linear strings. In contrast, tokenization divides these linearized strings into smaller chunks, or tokens, for additional processing. Tokenization is necessary for interpreting molecular structure representations, even if linearization is a natural part of the representation process. Previous studies, including the molecular graph generation work by Jin et al. (2018) [1], emphasize how crucial tokenization and linearization are to allowing efficient molecular structure modelling. Partitioning linear molecular representations can change how molecules are perceived because they are algorithmic abstractions. Any logical division of molecular structures based on SMILES strings is referred to as tokenization in this context.

The purpose of this overview of the literature is to put tokenization methods, denoisers, assessment metrics, and training approaches in the context of AI-driven molecular design. These elements are vital in determining how effective and efficient molecular production procedures are, which in turn affects how well drug development initiatives work. The review will examine the ways in which several subfields within this domain—such as model architecture, assessment techniques, and molecular representation—intersect and advance the discipline.

Tokenization in NLP

Textual data in NLP has historically been divided into "words" and "sentences" because of linguistic reasons and technological limitations. The large-scale components, or "sentences," are frequently divided into tiny units and regarded as separate entities from one another. Approximation and compromise have always been involved in the definition of these tiny units. These units must be linguistically motivated since they are subjected to linguistic annotations (such as part-of-speech tags, morphosyntactic annotation, and syntactic dependence information). However, a wide range of events make it extremely difficult to recognize and even define linguistic units, which are represented as word-forms by the Morphological Annotation Framework (MAF) ISO standard. Clément et al. (2005). [2]

Tokenization Techniques

Byte Pair Encoding

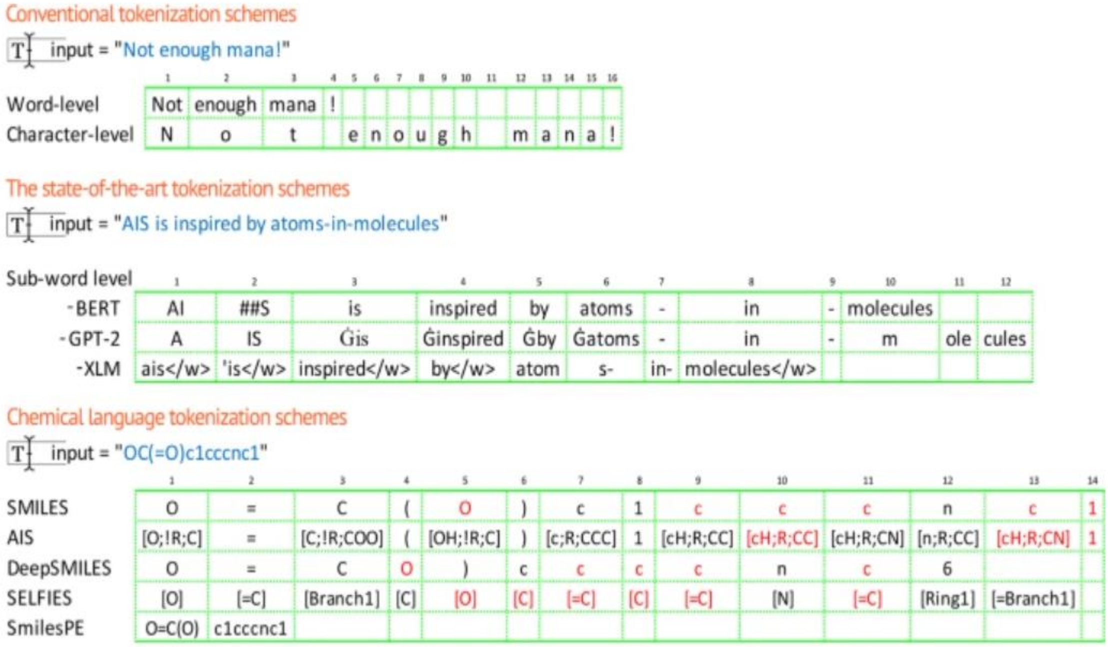
The sub word tokenization process known as "Byte Pair Encoding," or BPE for short, was first used in data compression schemes. But the NLP community quickly realized that it could be used to deconstruct text into more manageable, smaller chunks that a machine could comprehend and process. It is a straightforward data compression method that repeatedly swaps out the most common pair of bytes in a sequence for an empty byte. This algorithm is modified for word segmentation. Rather than combining frequently occurring byte pairs, we combine characters or character groups. Sennrich et al. (2016) [3].

BPE uses a data-driven approach to tokenization, in contrast to traditional approaches that frequently rely on predetermined vocabulary, or rules based on whitespace and punctuation marks. Iteratively combining frequently occurring neighbouring letters or character sequences (bigrams) into new symbols produces a vocabulary of multi-character units or sub words.

Tokenization is a fundamental aspect of preparing molecular data for input into AI models. Various techniques, such as SMILES, SELFIES, and SAFE strings, have been proposed for encoding molecular structures as sequences of symbols. Popova et al. (2018) [3] demonstrated the efficacy of deep reinforcement learning for de novo drug design, although the paper does not explicitly focus on tokenization. While tokenization techniques are essential for representing molecular structures, further research is needed to explore their impact on model performance and generalizability.

The development of natural language processing has given rise to sophisticated tokenization systems. While traditional tokenization techniques broke sentences down into words or characters, Figure 1 illustrates how state-of-the-art tokenization schemes, used in different models such as BERT, GPT-2, and XLM, divide words into sub-words to capture contextual links between them. Atom-wise tokenization of SMILES is widely utilized in cheminformatics to train chemical language models. New molecular representations like SELFIES and DeepSMILES, as well as specific tokenization systems like SmilesPE that mimic byte-pair encoding, have been introduced in addition to atom-wise SMILES tokenization [9].

Fig. 1



Conventional tokenization schemes

input = "Not enough mana!"

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Word-level	Not	enough	mana	!												
Character-level	N	o	t		e	n	o	u	g	h		m	a	n	a	!

The state-of-the-art tokenization schemes

input = "AIS is inspired by atoms-in-molecules"

Sub-word level	1	2	3	4	5	6	7	8	9	10	11	12
-BERT	AI	##S	is	inspired	by	atoms	-	in	-	molecules		
-GPT-2	A	IS	Ġis	Ġinspired	Ġby	Ġatoms	-	in	-	m	ole	cules
-XLM	ais</w>	'is</w>	inspired</w>	by</w>	atom	s-	in-	molecules</w>				

Chemical language tokenization schemes

input = "OC(=O)c1cccnc1"

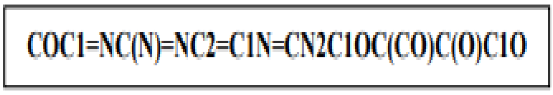
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
SMILES	O	=	C	(	O	)	c	1	c	c	c	n	c	1
AIS	[O;!R;C]	=	[C;!R;COO]	(	[OH;!R;C]	)	[c;R;CCC]	1	[cH;R;CC]	[cH;R;CC]	[cH;R;CN]	[n;R;CC]	[cH;R;CN]	1
DeepSMILES	O	=	C	O	)	c	c	c	c	n	c	6		
SELFIES	[O]	[=C]	[Branch1]	[C]	[O]	[C]	[=C]	[C]	[=C]	[N]	[=C]	[Ring1]	[=Branch1]	
SmilesPE	O=C(O)	c1cccnc1												

Comparison of conventional and modern tokenization schemes in NLP and the tokenization methods in the chemical language domain

SMILES Molecular Representation

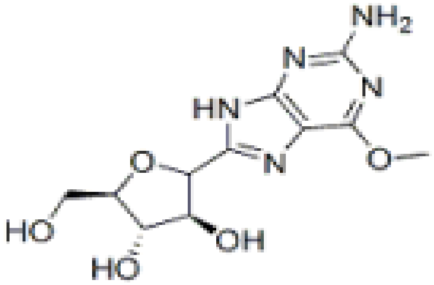
Simplified Molecular Input Line Entry, SMILES, was proposed by Weininger et al. It uses ASCII strings to describe the 2-D or 3-D structure of the molecule. It achieves this by using the depth-first tree traversal of the molecule graph. Below is an example of this [4].

Nelarabine – a chemotherapy drug used to treat leukaemia – provides an example of a linearized structure. Its molecular formula is C11H15N5O5. The following figure displays a SMILES string for this substance. [4]



COC1=NC(N)=NC2=C1N=CN2C1OC(CO)C(O)C1O

Nelarabine SMILES String – "A linearized string representation of Nelarabine whose two-dimensional structure is seen in the figure above. Here one can count e.g. 11 carbon atoms C and 5 nitrogen N and 5 oxygen O atoms. Hydrogen atoms H are implicit." [4]



Nelarabine 2-D Structure – "This formula displays one hexagonal and two pentagonal rings. Also seen atoms and groups (say NH2=Amino), linked by chemical bonds. In contrast to the first figure, some hydrogen atoms H are here explicit, while carbon atoms C are implied by unnamed vertices of the polygons." [4]

Related Works

BERT, introduced by Devlin et al. (2019) [9], marks a significant advancement in deep learning-based language models by employing the transformer architecture. Next sentence prediction (NSP) and masked language modelling (MLM) are the two primary pre-training techniques used by BERT. In the MLM task, BERT uses a "MASK" token to mask 15% of the words in input sequences, and then uses the semantic relationships between the masked and unmasked words to predict the original values of those words. This procedure entails using SoftMax to calculate the likelihood of every word in the vocabulary and appending a classification layer on top of the encoder output. The cross-entropy loss function of the model ignores the unmasked values and only concentrates on predicting the masked ones. Contrarily, NSP uses a binary classification task in which BERT must determine whether a given pair of sentences' second sentence comes first or not. BERT has proven to be exceptionally proficient at a variety of activities, including as language comprehension and question-answering.

Using the BERT framework and SMILES notations for molecular input, "MolBert" is a specific language model designed for chemical applications, as suggested by Fabian et al. (2020). MolBert is pre-trained using 1.6 million molecules from the Guacamole benchmark dataset, which is obtained from ChEMBL. The model is refined and tested on benchmarking tasks from MoleculeNet, including regression tasks like ESOL, FreeSolv, and Lipo datasets, as well as classification tasks like BACE, BBBP, and HIV datasets. Three self-supervised tasks are used to evaluate MolBert: (i) masked language modelling; (ii) predicting equivalence between two given SMILES inputs, where the second SMILES can be randomly selected from training data or a synonymous permutation; and (iii) predicting physicochemical properties of given molecules. [18]

Mixture of Denoisers

The integration of denoising strategies, such as mixture of denoisers (MoD) architectures, has shown promise in enhancing the robustness and diversity of generated molecules.

R-Denoiser

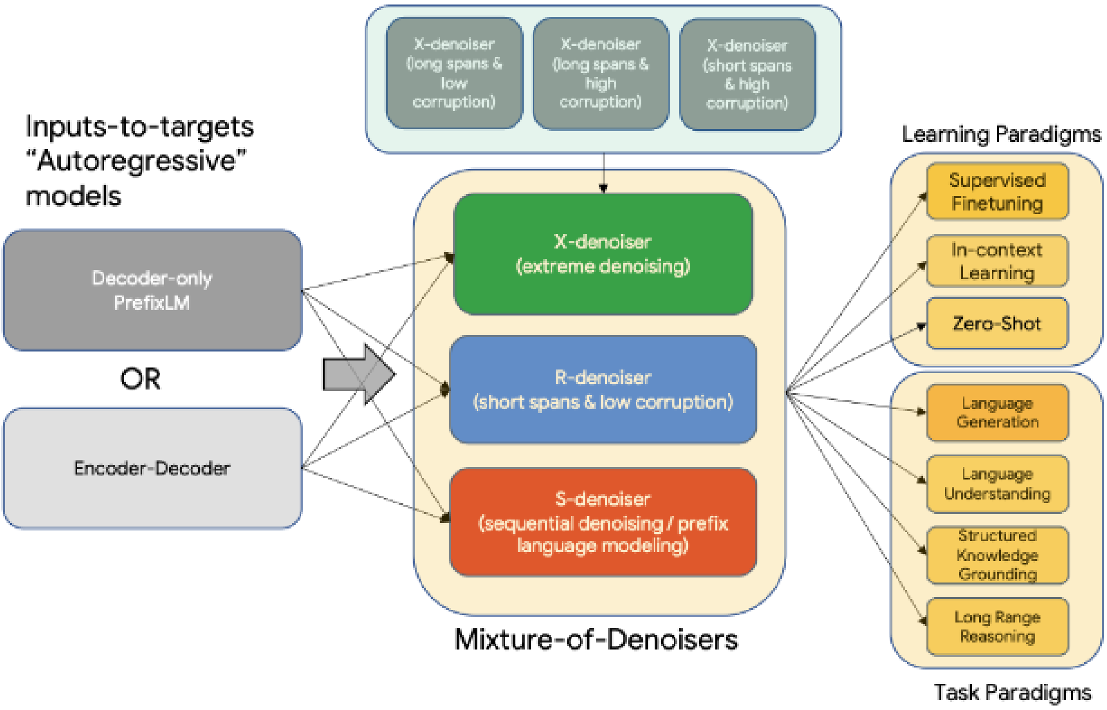
The regular denoising, which masks around 15% of the input tokens, is the standard span corruption first introduced in Raffel et al. (2019) [17]. It employs a range of 2 to 5 tokens as the span length. These brief periods may be more beneficial for learning than for developing writing fluency.

S-Denoiser:

A particular kind of denoising in which the inputs-to-targets work (prefix language modelling) is framed in a rigorous sequential manner. To achieve this, we simply divide the input sequence into two token sub sequences, context and target, so that the targets are independent on future knowledge. This is not the same as conventional span corruption, where a target token may be present earlier in time than a context token. Keep in mind that the context (prefix) maintains a bidirectional receptive field, just like in the Prefix-LM arrangement. We observe that S-Denoising that has little, or no memory is comparable to conventional causal language modelling. [19].

X-Denoiser

An extreme form of denoising that requires the model to recover a significant portion of the input, given a small to moderate portion of it. This mimics a scenario in which a model must produce a long goal from a memory containing a small amount of data. To do this, we choose to include cases with aggressive denoising, in which roughly half of the input sequence is obscured. By doing so, the span length and/or corruption rate are extended. A pre-training assignment is deemed intense if it spans a long time (e.g., more than 12 tokens) or has a high rate of corruption (e.g., more than 30%). X-denoising derives its motivation from its ability to interpolate between regular span corruption and aims akin to language models.


Inputs-to-targets
"Autoregressive"
models
Decoder-only
PrefixLM
OR
Encoder-Decoder
X-denoiser (long spans & low corruption)
X-denoiser (long spans & high corruption)
X-denoiser (short spans & high corruption)
X-denoiser (extreme denoising)
R-denoiser (short spans & low corruption)
S-denoiser (sequential denoising / prefix language modeling)
Mixture-of-Denoisers
Learning Paradigms
Supervised Finetuning
In-context Learning
Zero-Shot
Language Generation
Language Understanding
Structured Knowledge Grounding
Long Range Reasoning
Task Paradigms

UL2 framework, source: https://huggingface.co/google/ul2

Evaluation Metrics

Evaluating the quality and utility of generated molecules requires the adoption of comprehensive evaluation metrics. Schwaller et al. (2019) [12] proposed a neural sequence-to-sequence model for predicting outcomes of complex organic chemistry reactions, emphasizing the importance of diverse evaluation metrics such as molecular validity and synthetic accessibility. While existing literature provides insights into individual metrics, there is a need for standardized evaluation frameworks that encompass multiple aspects of molecular quality.

Validity

The proportion of SMILES that are produced and able to be transformed into legitimate compounds. The degree to which a generative model has mastered the SMILES syntax and the chemical principles involved in building molecules is measured by its validity.

Novelty

Novelty is defined as the proportion of legitimate molecules absent from the training set. A low novelty could mean that the training set was overfitted by the generative model.

Uniqueness

The proportion of legitimate molecules with a unique composition. A low uniqueness would suggest mode collapse, when the model sample a small subset of molecules from a small number of distinct regions of the chemical universe.

Internal Diversity

An array of produced molecules' chemical diversity is measured by its internal diversity. Another reliable sign of mode collapse is internal variety.

$$\text{Internal diversity} = 1 - \frac{1}{|G|} \sum_{(x_1,x_1)\in G\times G} T(x_1, x_2)$$

A representation of the internal diversity of a generated set of molecules. A larger value means a higher chemical diversity [16].

Training Strategies

Optimizing training strategies is essential for enhancing the efficiency and effectiveness of AI-driven molecular design. Segler et al. (2018) introduced a planning framework for chemical syntheses using deep neural networks and symbolic AI, underscoring the significance of training methodologies in guiding molecular generation processes. However, the scalability and generalizability of training strategies remain areas of ongoing research, necessitating further exploration to address challenges associated with dataset size and computational resources.

Sequence-based tactics are emerging as an alternative to graph-based approaches. The basic idea is to use a text representation of the reactants, reagents, and products (typically a simplified molecular-input line-entry system, or SMILES) to consider reaction prediction as machine translation from one language (reactants–reagents) to another (products). Schwaller et al. [12] have shown that sequence-to-sequence models (seq-2-seq) that make analogies between organic chemistry and human language could compete with graph-based methods. In the two previous seq-2-seq attempts, the encoder and decoder were based on recurrent neural networks (RNNs), with a single single-head attention layer placed in between.

Seq2seq models in organic reaction prediction and retrosynthesis

Retrosynthesis is the opposite of reaction prediction [13]. The objective is to identify potential reactants given a product molecule. Retrosynthesis allows for the possibility of more than one target string being right, unlike major product prediction. For example, a product may result from two distinct pairs of reactants. When a seq2seq model has no clear aim, training it can be more challenging. Liu et al [13]. made the first attempt to use a seq2seq model in retrosynthesis. They used a set of 50 000 reactions extracted and curated by Schneider et al. Ten distinct reaction classes comprise the reactions from that set, which also includes stereochemical data. All things considered; no prior study was able to fully illustrate the potential of seq2seq models.

Conclusions

This research review concludes by highlighting the significance of training strategies, evaluation metrics, tokenization approaches, and denoisers in AI-driven molecular design. Even while each of these areas has seen tremendous advancements, there are still gaps in our knowledge of how they all affect model performance and how well it applies to actual drug discovery situations. For the discipline to advance and AI in molecular design to reach its full potential, interdisciplinary cooperation and ongoing research are essential.

Way Forward for the Project

Building upon the insights gained from this literature review, the next steps for the project involve conducting empirical studies to systematically evaluate the interplay between different components of AI-driven molecular design. By developing experimental frameworks that incorporate diverse tokenization techniques, denoising strategies, evaluation metrics, and training methodologies, the project aims to contribute to the optimization of molecular generation processes and facilitate the discovery of novel therapeutic molecules.

References

[1] Wengong Jin, Kevin Yang, Regina Barzilay, Tommi Jaakkola (2018) Learning Multimodal Graph-to-Graph Translation for Molecular Optimization. https://doi.org/10.48550/arXiv.1812.01070

[2] Lionel Clément, Eric De la Clergerie, and Lionel Net. (2005). Maf: a morphosyntactic annotation framework.

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch. (2016). Neural Machine Translation of Rare Words with Subword Units. https://aclanthology.org/P16-1162

[4] Weininger, D., Weininger, A., and Weininger, J.L. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation, J. Chem. Inf. Comput. Sci, 29, pp. 97-101

[5] Mariya Popova et al. (2018), Deep reinforcement learning for de novo drug design.Sci.Adv.4, eaap7885.DOI:10.1126/sciadv.aap7885

[6] Exman, Iaakov. (2014). Web Search of New Linearized Medical Drug Leads. 10.5220/0003705401080115.

[7] Ucak, U.V., Ashyrmamatov, I. & Lee, J. (2023) Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. J Cheminform 15, 55. https://doi.org/10.1186/s13321-023-00725-9

[8] Domingo M, Garcıa-Martınez M, Helle A, et al (2018) How Much Does Tokenization Affect Neural Machine Translation? Arxiv.https://doi.org/10.48550/arxiv.1812.08621

[9] Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv.https://doi.org/10.48550/arXiv.1810.04805

[10] Radford A, Wu J, Child R, Luan D, Amodei D & Sutskever I (2019) Language Models are Unsupervised Multitask Learners. OpenAI.

https://www.openai.com/blog/better-language-models/

[11] Lample G, Conneau A (2019) Cross-lingual language model pretraining. arXiv. https://doi.org/10.48550/arXiv.1901.07291

[12] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. (2019) Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction ACS Central Science, 1572-1583 DOI: 10.1021/acscentsci.9b00576

[13] Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J, Luu Nguyen Q, Ho S, Sloane J, Wender P, Pande V. (2017) Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. ACS Cent Sci. 1103-1113. doi: 10.1021/acscentsci.7b00303. Epub 2017 Sep 5. PMID: 29104927; PMCID: PMC5658761.

[14] Schneider N, Stiefl N, Landrum GA. (2016). What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. J Chem Inf Model.:2336-2346. doi: 10.1021/acs.jcim.6b00564.

[15] M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, (2017). Molecular de novo design through deep reinforcement
learning, http://arxiv.org/abs/1704.07555.

[16] Xinhao Li and Denis Fourches. (2021) SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

[18] Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, Tunca Doğan (2023) SELFormer: Molecular Representation Learning via SELFIES Language Models. https://arxiv.org/pdf/2304.04662.pdf

[19] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, Donald Metzler. (2022). Unifying Language Learning Paradigms. https://doi.org/10.48550/arXiv.2205.05131