Journal of Cheminformatics

RESEARCH

Open Access

Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization

Umit V. Ucak[1†], Islambek Ashyrmamatov[2†] and Juyong Lee[3*]

Abstract

Tokenization is an important preprocessing step in natural language processing that may have a significant influence on prediction quality. This research showed that the traditional SMILES tokenization has a certain limitation that results in tokens failing to reflect the true nature of molecules. To address this issue, we developed the atom-in-SMILES tokenization scheme that eliminates ambiguities in the generic nature of SMILES tokens. Our results in multiple chemical translation and molecular property prediction tasks demonstrate that proper tokenization has a significant impact on prediction quality. In terms of prediction accuracy and token degeneration, atom-in-SMILES is more effective method in generating higher-quality SMILES sequences from AI-based chemical models compared to other tokenization and representation schemes. We investigated the degrees of token degeneration of various schemes and analyzed their adverse effects on prediction quality. Additionally, token-level repetitions were quantified, and generated examples were incorporated for qualitative examination. We believe that the atom-in-SMILES tokenization has a great potential to be adopted by broad related scientific communities, as it provides chemically accurate, tailor-made tokens for molecular property prediction, chemical translation, and molecular generative models.

Keywords Atom-in-SMILES, Tokenization, Repetition, Chemical language processing

Introduction

Tokenization is an essential preprocessing step for sequential data to train and use natural language processing (NLP) models. However, insufficient attention has been devoted to its effects on chemical applications. Tokenization can significantly influence prediction quality within the framework of text generation [1]. In the field of chemistry, it covers processes used to split linear molecular representations into their constituent elements. As linear molecular representations are algorithmic abstractions, their partitioning can alter the perception of molecules. Herein, tokenization refers to any logical partitioning of molecular structures based on SMILES strings.

In general, a molecule can be perceived as an inherent whole, owing to the internal relationships among its atomic components. Simplified Molecular Input Line Entry System (SMILES) strings [2], the most commonly used molecular representation, are also defined to be meaningful as a whole. They represent molecular objects, which are rigid bodies and completely different from their constituent atoms. Notably, any sensible partitioning of a molecule will produce meaningful fragments.

†Umit V. Ucak and Islambek Ashyrmamatov have contributed equally to this work.

*Correspondence:
Juyong Lee
nicole23@snu.ac.kr
[1] Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea
[2] College of Pharmacy, Seoul National University, Seoul, Republic of Korea
[3] Research Institute of Pharmaceutical Science, Seoul National University, Seoul, Republic of Korea

However, to some extent, the abstraction level of the process will get tangled because considering atoms in a molecule is not a "realistic" approach. Likewise, SMILES tokens are merely sensible linear cuts of a string with reduced dimensionality.

A typical tokenization of SMILES is performed atom-wise, i.e., character-wise. However, the SMILES representation consists of a small number of distinct characters including atomic symbols, integers for ring closure, and special symbols for bonds and chirality. In other words, all atoms with the same atomic number are represented identically. However, the characteristics of each atom even with the same atomic number may differ significantly based on its environment. This statement is similar to the concept of atoms in molecules (AIM) [3], which describes the nature of molecules based on the electron density distribution attracted by each atom. Thus, the conventional atom-wise tokenization of SMILES may be too abstract and chemically inaccurate and it may obscure the learning process of a model and the understanding of the results.

Herein, we point out an analogy between the natural (spoken) language and constructed language of chemistry (see Table 1). The analogy provides a motivational ground for the use of NLP methods in chemistry problems [4]. The intended analogy relies on the part-whole relationship [5, 6] and suggests that molecular substructures (typically composed of several atoms) can be considered as chemical "words" for the linguistic treatment of chemical language. However, in practice, chemical words often become the tokens of SMILES that consists of atomic symbols and characters representing topological characteristics, such as ring-closure or branches, which do not correspond to physical atoms. In this context, atoms are present in molecules by neglecting an essential aspect of the chemical reality. In the following paragraph, we analyze the token characteristics of sentences and SMILES strings for insight into the influence of the latter format on the translation mechanic.

According to the sentence length distribution of various language corpora, a well-written sentence contains 15–20 words on average [7]. The average sequence length of a SMILES string is typically three times longer than a natural language, whereas the token space is at least 1000 times smaller than any developed language [8]. This is a consequence of repetitive tokens observed in SMILES strings. The most distinguishing feature of SMILES representation is the token repeat, which causes atoms of molecules to be indistinguishable in the token space. The repetitive nature of SMILES syntax adds to the more general issue of neural machine translation (NMT) decoders, yielding degenerative outcomes [9, 10].

Token order is another aspect of this comparison. Although the order of words in a sentence can be altered to enhance tone, meaning, or fluency, this cannot be applied to molecules. In fact, a single molecule can equally be represented by hundreds of SMILES enumerations depending on its topology (more if branches and cyclic fragments exist) [11]. Canonical SMILES refers to one of those many allowed permutations obtained by a unique and consistent atom numbering. In essence, while words tend to retain their semantic significance as they transition from isolated to contextual settings, with only minor semantic shifts that may occur over time, the same does not hold true for atoms within the realm of chemistry. For example, the atomic symbols within a SMILES string are treated equivalently to those in isolation, signifying that the chemical significance of these symbols is upheld throughout the tokenization process. Thus, tokens such as carbon (C) may appear identical in different molecules despite their actual differences in chemical composition. However, it is important to distinguish between atom-in-SMILES (AIS, analogous to AIM) and corresponding tokens, as atoms lose their identities when they form molecules.

Inspired by the aforementioned comparative analysis outlined in Table 1, we develop a tokenization framework by introducing environmental information and show that it corroborates the chemical viewpoint. In recent decades, various methods have been developed to enhance or extend the SMILES language. Few of these methods include BigSMILES for describing macromolecules [12], CurlySMILES for supramolecular structures and nanodevices [13], CXSMILES for storing the special features of molecules [14], OpenSMILES specification for specifying the stereochemistry and chirality [15], DeepSMILES and SELFIES for machine learning applications [16, 17], and canonicalization algorithms [18, 19]. The aforementioned approaches effectively solve particular problems originating from the internal structure of SMILES. In our approach, we do not treat syntactic problems; rather, we redefine SMILES tokens by introducing environmental information. To consider local chemical environments, atom environments (AEs) are used, which are circular atom-centered topological molecular fragments created with predefined radii of covalent bonds. Hence, our

Table 1 Comparison of the important aspects of natural and chemical languages within the NLP framework

Aspects	Natural language	SMILES language
Sequence length	15-20 words	~ 3 times higher
Token space	>100K	~ 1000 times smaller
Token order	Tone, meaning, fluency	${}_nC_2$ alternatives*
Meaning-wise	isolation ≡ context	isolation ≡ context

*practically less due to the rules of chemistry

approach entails utilizing AEs to produce environment-aware atomic tokens analogous to atom-in-molecules. We term this custom tokenization scheme Atom-in-SMILES, AIS.

The AIS tokenization integrates key aspects of SMILES and AEs [20]. The proposed approach accommodates all relevant information for a seamless bidirectional transformation between the two representations, to ensure practical implementation. We demonstrate that the AIS scheme performs better in translating between equivalent string representations of molecules using an exceptionally challenging dataset. It was also observed that training with AIS tokenization leads to more accurate models for single-step retrosynthetic pathway prediction, and molecular property prediction tasks. We also evaluated prediction qualities by comparing AIS tokenization with the existing schemes: canonical SMILES-based tokens, atom-wise and SMILES pair encoding (SmilesPE) [21], SELFIES, and DeepSMILES tokens. We show that AIS tokenization reflects the true chemical context, delivers better performance, and reduces token degeneration by 10%.

Implementation

Advanced tokenization schemes have emerged as a result of the evolution of natural language processing. Figure 1 shows that state-of-the-art tokenization schemes, like BERT [22], GPT-2 [23], and XLM [24], divide words into sub-words to capture contextual relationships between them while conventional tokenization schemes used to break down sentences into words or characters. In the field of cheminformatics, atom-wise tokenization of SMILES is primarily used for training chemical language models. In addition to atom-wise SMILES tokenization, new molecular representations have been introduced such as SELFIES and DeepSMILES, and specialized tokenization schemes like SmilesPE imitating byte-pair encoding.

In the token space generated by atom-wise SMILES tokenization, all atoms with the identical atomic numbers are indistinguishable. As a toy example, in a glycine molecule (Fig. 2) carbons are represented as two identical carbon atoms following tokenization. Oxygen atoms are also treated similarly. Hypothetical atomic constituents obtained by tokenization are often degenerated. This is an intrinsic feature of SMILES representation, which does not correspond to chemical reality.

We propose the AIS tokenization scheme that expresses. The most natural formulation of this proposition is as follows. Let T_1 and T_2 be the token spaces of SMILES and AIS, respectively, and $f : T_1 \rightarrow T_2$ be a mapping, which is one-to-one and onto; then,

$$f(X) = \begin{cases} AE|_{X_{\text{central}}} & \text{if X is an atom} \\ X & \text{otherwise.} \end{cases} \quad (1)$$

For any SMILES string, the function f simply tweaks each atom by selecting it as the central atom of the

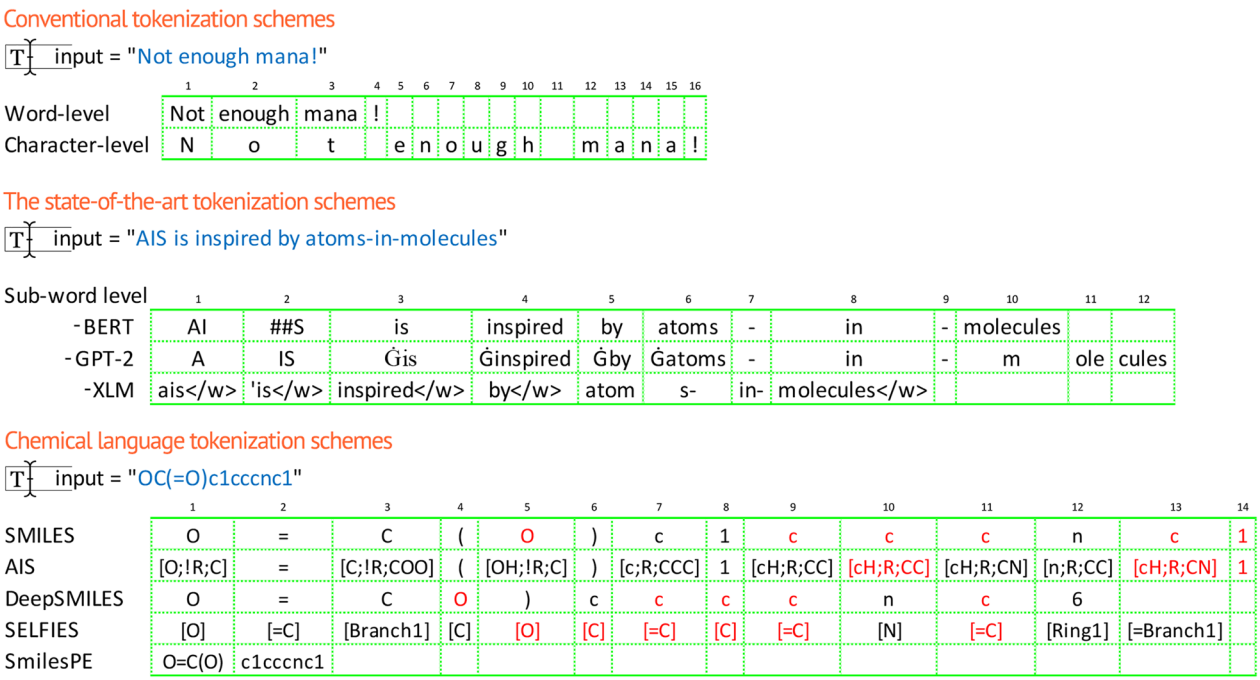

Fig. 1 Comparison of conventional and modern tokenization schemes in NLP and the tokenization methods in the chemical language domain

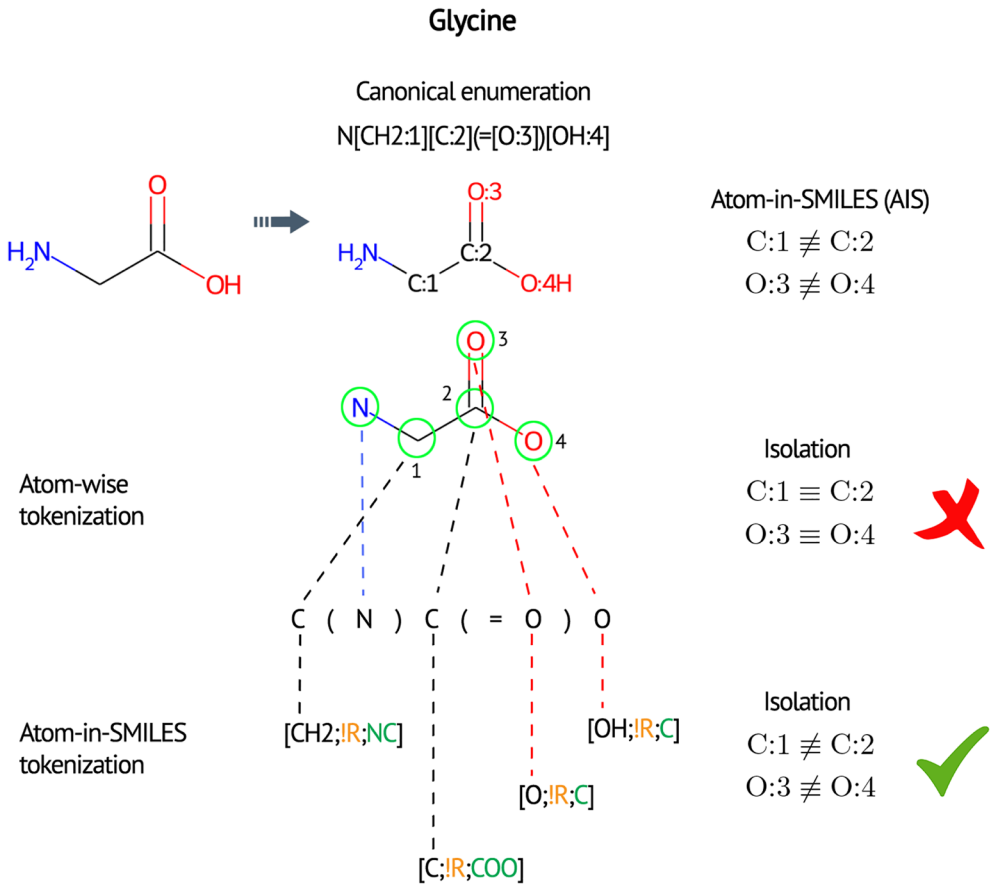

Fig. 2 A toy example illustrating the major differences between AIS and conventional SMILES tokenizations. The formal description of AIS tokenization contains three primary elements, (i) central atom, (ii) ring information, and (iii) neighbor atoms information, interacting with the central atom. The formalism ties everything together within a square bracket separated by a semi-colon. The chirality information can be attached to the central atom, which is labeled with either @ or @ @ suffixes. Aromaticity is reflected on the central atom with a lower case letter. Hydrogen atoms are explicitly specified on central atoms. The hybridization and bonding nature of organic elements can be easily deduced

corresponding atomic environments (AEs); otherwise, it is an identity operator from the token space. The chirality and aromaticity information of the central atom are preserved through the above-described mapping. Bond order and hybridization are the two intrinsic dimensions of AIS tokens. As f is invertible, SMILES strings can be fully recovered by the SMILES projection. The proposed algorithm for generating AIS tokenization can be described through the presented pseudo-code (Algorithm 1). The algorithm works by iterating over the atoms in a SMILES string and generates rich, environment-aware variants.

Introducing the neighboring atoms interacting with the central atom generates tokens with greater diversity. As shown in Fig. 2 carbon and oxygen atoms are well distinguished according to their local chemical environments (C:1 ≠ C:2, O:3 ≠ O:4). The token space stretches relative to the atom-wise tokenization. As another example, shown in Fig. 3, we can consider the tokens of the following aromatic molecules: `Clc1csc2nc3ncncn3c12` and `Clc1csc2nc3nnccn3c12`. The atom-wise tokens of these molecules are identical. However, the set of the symmetric difference of AIS tokens has three members, `[cH;R;CN]`, `[cH;R;NN]`, and `[n;R;CN]`, rendering the carbon-nitrogen swap recognizable.

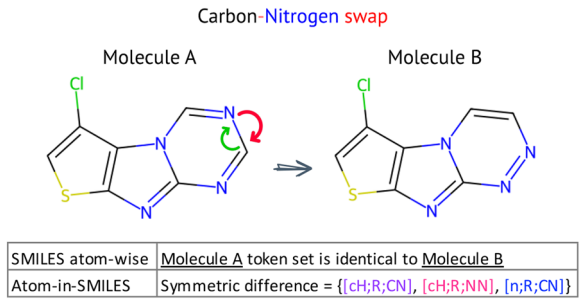

Fig. 3 Token set comparison of two highly similar molecules. The molecules differ only in the position of a carbon and nitrogen atom in one of the rings

Algorithm 1 Atom-in-SMILES tokenization formalism

```
function AtomInSmiles(smiles)
    if (∃ atomMapNums in smiles) then
        mol ← smiles
    else
        mol ← canonical ← smiles
    end if
    D ← {}; atoms ← {}
    for all atom ∈ mol do
        atoms ← GetSmarts()
        atomId ← GetAtomMapNum()
        sym ← GetSymbol()
        if atom is aromatic then sym ← sym.lower()
        if (∃ charge) then sym ← sym + GetFormalCharge()
        if (∃ chirality) then sym ← sym + GetChiralTag()
        sym ← sym + GetTotalNumHs()
        if atom is in a ring then ring ← 'R' else ring ← '!R'
        neighbs ← GetNeighbors()
        D[atomId] ← '[Sym;Ring;Neighbs]'
    end for
    AIS ← {}
    tokens ← Tokenize(smiles)
    for all token ∈ tokens do
        if token ∈ atoms then
            atomId ← GetAtomMapNum(token)
            AIS ← D[atomId]
        else
            AIS ← token
        end if
    end for
    return AIS
end function
```

Figure 4 provides insight into the inherent properties of various molecular representations, revealing their expressive power, token diversity, and chemical relevance. We evaluated the distributions of tokens and normalized repetition rates across a diverse set of molecular datasets with a wide range of structural complexities and configurational changes, such as coordination compounds and ligands (metal complexes from Crystallography Open Database [25]), ring structures and functional groups (steroids [26]), long-chain formations (phospholipids and ionizable lipids [27]), complex and diverse structures (natural products [28]), small organic molecules (drugs [29]), and configurational changes in molecular structure (octane isomers). Single-token repetition can be easily quantified as rep-1 $= \sum_{t=1}^{|s|}[s_t \in s_{t-w-1:t-1}]$, where s and $|s|$ denote the prediction and token count respectively [10]. We kept the number of considered previous token w sufficiently large (as large as the maximum sequence length). Normalized repetition rates, which measure the ratio of single-token repetitions to sequence length, is used to provide a meaningful measure of expressiveness. Lower repetition rates indicate more diverse and informative token sets that can alleviate the problem of degeneracy observed in model outcomes.

In Fig. 4, AIS tokens exhibit consistently lower repetition rates compared to SMILES, SELFIES, and DeepSMILES, indicating a higher level of expressiveness. This difference in expressive power is particularly evident in drugs, natural products, and steroid datasets. However, in expressing long chains, as in the case of lipids, all tokenization schemes struggle. One limitation of AIS is that it lacks the ability to distinguish environmentally similar substructures or those with a symmetry plane since it only considers nearest neighborhoods. The SmilesPE representation exhibits a low-lying distribution due to the relatively low number of pseudo-substructures with fewer or zero repetitions. It is worth noting that inherent repetitions in molecular representations can exacerbate the repetition problem observed in NLP model outcomes, highlighting the importance of diverse and informative token sets.

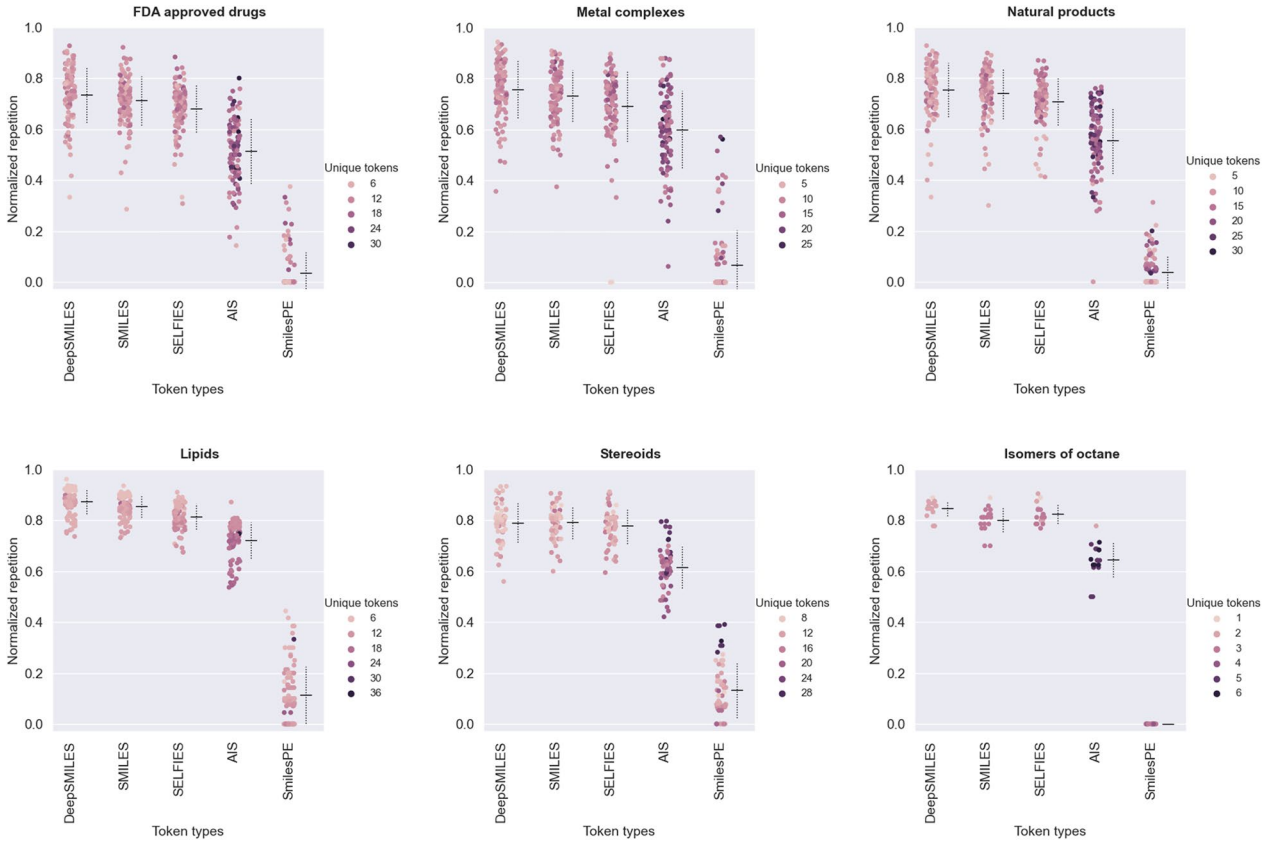

Fig. 4 Comparison of expressiveness and normalized repetition rates across various molecular representations. Distributions showcasing the distinct characteristics of tokenization schemes on representative datasets, each designed to test different facets of molecular structures such as coordination compounds, ligands (metal complexes), ring structures and functional groups (steroids), long-chain formations (phospholipids, ionizable lipids), complex and diverse structures (natural products), small organic molecules (drugs), and configurational changes in molecular structure (octane isomers). Each dataset contains one hundred members, with the exception of steroids (59 members) and octane isomers (18 members). The mean values of normalized repetitions and deviations from the mean are visually represented as horizontal and dashed vertical lines, respectively, accompanying the distributions

Results and discussion

We tested the AIS tokenization on three challenging tasks: (i) input–output equivalent mapping (SMILES canonicalization), (ii) single-step retrosynthetic prediction, and iii) molecular property prediction. For the first two functionality tests, we utilized NMT framework that translates sequences from the source to the target domain with the most promising attention-based transformer encoder-decoder architecture [30, 31]. We trained our models for 200,000 steps with the Adam optimizer, negative log-likelihood loss, and cyclic learning rate scheduler. For these tasks, we report the percentage of exact prediction.

Input–output equivalent mapping

First, we tested how the learning efficiency of an NMT model is affected by the choice of the tokenization scheme, on the task of converting non-canonical SMILES strings into their canonical form. For rigorous test, we generated extremely confusing datasets consisting of many similar strings. To generate the datasets, we used the predefined subsets of the GDB-13 [32] database that contains drug-like molecules with up to 13 heavy atoms which consist of C, N, O, S, and Cl. The subsets were generated by applying cumulative pre-defined constraints [33, 34], which were named as follows: `a`: No cyclic HetHet Bond; `b`: No acyclic HetHet Bond; `c`: Stable FG; `d`: No cyclic C=C and C:C bonds; `e`: No acyclic C=C and C:C bonds; `f`: No small rings; `g`: Fragment-like, and `h`: Scaffold-like. Our training dataset consisted of one million randomly sampled molecules taken from the GDB-13, combined with 150K randomly sampled from the most stringent GDB-13 subset `abcdefgh`. We augmented the subset at different levels (×10, ×30, and ×50) to make the training set more confusing. This approach resulted in training datasets with a high degree of similarity between the input (non-canonical instances)

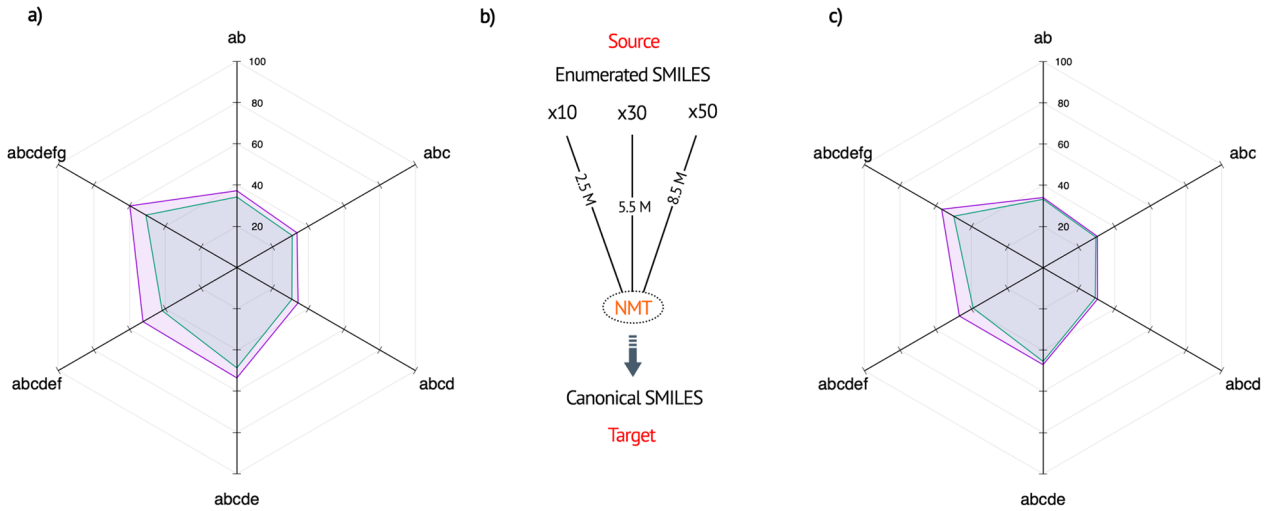

Fig. 5 Performance of atom-wise (blue) and atom-in-SMILES (purple) tokenization schemes tested on various restricted GDB-13 test sets [33]. **a** Test results of ×10 augmented training set. **b** Model overview. **c** Test results of ×50 augmented training set. The training is conducted with one million randomly sampled molecules taken from the GDB-13, combined with 150K randomly sampled subset of the strictest cumulative `abcdefgh` data, which we augmented at different levels (×10, ×30, and ×50)

Table 2 Performance of atom-wise and atom-in-SMILES tokenization schemes tested on various restricted GDB-13 test sets [33]

GDB-13 subsets [33] (cumulative)	Prediction accuracy (%)					
	Atom-wise			Atom-in-SMILES		
	x10	x30	x50	x10	x30	x50
`ab`	34.2	34.3	33.2	37.3	35.9	34.1
`abc`	31.0	30.8	29.6	33.7	32.1	30.4
`abcd`	30.8	30.4	29.2	34.3	32.3	30.5
`abcde`	48.7	47.6	45.5	53.6	50.0	47.0
`abcdef`	41.8	40.6	39.1	52.5	49.6	46.9
`abcdefg`	50.9	50.9	50.0	59.9	58.6	56.8

The training is conducted with one million randomly sampled molecules taken from the GDB-13, combined with 150K randomly sampled subset of the strictest cumulative `abcdefgh` data, which we augmented at different levels (×10, ×30, and ×50)

and output (only canonical enumerations) SMILES strings, making it difficult to discern variations.

We quantified the performance on large (20K) GDB-13 disjoint test sets of varying constraints (see Table 2). To highlight the benefits of our token design over SMILES tokenization, we utilized an approach shown in Fig. 5b. Table 2 and Fig. 5a, c demonstrate the limitations of SMILES tokenization and the characteristics of our tokens. The atom-in-SMILES scheme outperformed the SMILES atom-wise scheme on all subsets and augmentation levels, with increasing performance gaps for more restrictive subsets (more similar). The highest prediction accuracy of 59.9% (x10) and 56.8% (x50) was achieved on the subset `abcdefg`, compared to 50.9% (x10) and 50.0% (x50) for the atom-wise scheme.

In our experiments, we observed that the added complexity by data augmentation resulted in a degradation of performance, different from the typical degradation observed in overly complex models (overfitting) [35]. Atom-wise tokens struggled to handle the increasing complexity, resulting in a performance deficit of up to 10.7% on the `abcdef` subset. Notably, as the level of augmentation increased, the model's token-level probabilities decreased. However, we found that the AIS tokenization, trained on a dataset of extremely similar molecules, was better equipped to handle this problem. The greater string similarity led to consistent improvements in predictive power, which we attribute to the richer and more expressive representation of AIS tokens.

Single-step retrosynthesis and token degeneration

Retrosynthetic prediction is a challenging task in organic synthesis that involves breaking down a target molecule into precursor molecules using a set of reaction templates. This process helps chemists identify potential routes for synthesizing novel chemical structures. However, conventional template-based methods have limitations such as coverage and template generation issues [36, 37], and can be computationally expensive [38]. Additionally, atoms have not been successfully mapped between products and reactants in these methods [39]. To address these challenges, we implemented a template-free, direct translational method to suggest reactant candidates, which is extremely similar to the concept proposed by several groups [40–44]. These approaches can provide high-quality and complete recommendations without the need for hand-crafted templates [45] or pre-existing reaction databases [46, 47].

We adopted the open-source ca. USPTO-50K reaction benchmark dataset that is widely used for a single-step retrosynthesis prediction task. This dataset was a subset of a larger collection from the U.S. patent literature obtained with a text-mining approach [48, 49]. As a preprocessing step, we removed sequences longer than 150 tokens. The prediction quality is assessed by top-1 accuracy, string match. Additionally, we reported Tanimoto exactness (with hashed Morgan Fingerprint radius of 3 and bit size of 2048 [50]) since the predicted structures might fail on the string match tests [51], but still can map to correct ground truth due to multiplicity of SMILES representation. To determine the effect of the tokenization on the prediction quality, we compared the performance of the AIS tokenization with two other SMILES-based tokenization schemes, namely, atom-wise and SmilesPE, and two molecular representations DeepSMILES and SELFIES.

Repetition is a well-known issue in text generation models, where multiple tokens predict the same subsequent token with high probability [9, 56], leading to the generation of repetitive sequences. A sequence is said to have a repetition subsequence if and only if it contains at least two adjacent identical continuous subsequences [56]. Large-scale language models such as Transformer and GPT-2 have shown to exhibit this issue, resulting in a negative impact on the quality of generated text. The Table 3 demonstrates different types of token degeneration, including single-word repetition, phrase-level repetition, sentence-level repetition, structural repetition, and subsequential repetition. The examples are drawn from a range of NLP tasks, such as sentence completion, summarization, generation from an initial tag line, product review generation, protein sequence generation, and molecule captioning. This emphasizes the prevalence of token degeneration and highlights the importance of addressing this issue to ensure the generation of high-quality natural language text.

Herein, we observed that molecular prediction tasks are also susceptible to token repetition. With the careful examination of non-exact predictions, we were able to summarize the common forms of problematic outcomes. Figure 6 displays six typical examples of token repetition in SMILES predictions within an NMT retrosynthesis framework: long head and tail, repetitive rings and chains, and halogen repetitions on aliphatic and aromatic carbons. These outcomes are considered to be the most probable by the model and have a negative impact on the quality of predictions. The long head and tail result from the repeated addition of identical or similar substructures to a terminal, whereas repetitive rings and chains occur due to the repeated addition of the same substructure. The halogen repetitions on aliphatic and aromatic carbons occur when the model repeats the same halogen substitution on similar carbons. Understanding and addressing these problematic outcomes is crucial for the development of accurate molecular prediction models.

Methods for quantifying the propensity of subsequent repetition are adapted from the recent studies by Welleck [10] and Fu [56] on neural text degeneration. We focused more on token-level measure for repetition than sequence-level repetition [9] because NMT typically uses a maximum log-likelihood training objective that is concerned with optimizing next-token conditional distributions. We used a token-level measure for repetition, rep-l, that counts the single-token repeats appearing in the preceding tokens. As there are so many single-token repeats appear in the ground truth SMILES, we reported the number of predicted SMILES with repetition rate higher than the ground truth.

In Table 4, top-1 string exact and Tanimoto exact accuracy are listed for various tokenization schemes along with the number of predicted SMILES with repeated tokens, $\text{rep-l}|_P - \text{rep-l}|_{GT} \geq 2$, where P and GT refer to prediction and ground truth. We observed performance gains using the AIS tokenization, outperforming the baseline by 4.3% in string exacts and 2.9% in T_c exacts. Our methodology, having the fewest single-token repeats, alleviated the repetition problem by approximately 10% compared to the atom-wise tokenization scheme. DeepSMILES exhibited the worst degenerate repetition among all tokenization schemes, but its overall accuracy in predicting retrosynthesis was 3.5% lower than the baseline on average. Regardless of the repetition rate, SELFIES showed lower retrosynthesis prediction accuracy than the baseline of SMILES atom-wise tokenization. The overall performance of SmilesPE was about only half of the baseline. This clearly demonstrates that

Table 3 Degenerate repetition observed in various tasks from state-of-the-art large-scale language models

Task [model]	Context	[Repetition pattern] output
Sentence completion [GPT2] ref. [10]	**Prefix:** Lyrically the song has excerpts of different languages including French, Spanish,	[Single-word repetition] German, Italian, Portuguese, Spanish, **Portuguese**, Portuguese, Portuguese, Portuguese, Portuguese, Portuguese, Portuguese, Portuguese, Portuguese, .
Summarization [Transformer] ref. [52]	**Article**	[Phrase-level repetition] ...cristiano ronaldo's best footballer whereas barcelona talisman lionel messi's **best footballer** whereas barcelona's best footballer whereas barcelona talisman lionel lee whereas's best footballer's. (Truncated)
Generation from an initial tag line [GPT-2] ref. [9]	**Tag:** So what's new in my life? 09/11/18 - Just got back from vacation	[Sentence-level repetition] Just got back from vacation. 09/10/18 - **Just got back from vacation**. Just got back from vacation. 09/09/18 - Just got back from vacation. Just got back from vacation. 09/08/18 - Just got back from vacation. Just got back from vacation.
Product review generation [GPT-2] ref. [53]	**Initial context**	[Structural repetition] Great movie, although took a while to see at first it held my interest and kept me interested, plus i thought it was extremely good. also it was **very good**.
Protein sequence generation [ProtGPT-2] ref. [54]	**No context**	[Subsequential repetition] MSNDTPTHDPTPPAPAPAPE**PAPAPAPE**PAPAPAPEPAPAPAPEPAPAPAPEPAPAPAPEPAPAPAPEPAPAPAPEPAPAPAPEPAPAPAPEPAPAPAPE.
Molecule captioning [Transformer] ref. [55]	**SMILES:** `CC[N+](CC)=C1C=CC2=NC3=C(OC2=C1)C=C(N)C(C) =C3`	[Single-word repetition] the molecule is a deuterated compound that is **is** is is an isotopologue of chloroform in which the four hydrogen atoms have been replaced by deuterium. (Truncated)

The examples contain single-word repetitions, phrase-level repetitions, sentence-level repetitions, structural repetitions where tokens may vary within a repeating phrase, and subsequential repetitions. The first repeated unit in each example is emphasized in bold.

the pseudo-substructures obtained with SmilesPE were not sufficient to capture the chemical change with ease.

Molecular property prediction tasks

The Table 5 provides a comprehensive comparison of various tokenization schemes for molecular property prediction tasks, highlighting the significance of tokenization in this domain. To evaluate the performance of these schemes, we used the MoleculeNet benchmark datasets [58], which includes three regression tasks: ESOL for estimating solubility, FreeSolv for hydration free energies, and Lipophilicity for octanol/water distribution coefficient, logD at pH 7.4 and three binary classification tasks: BBBP for barrier permeability, BACE for predicting binding results for a set of inhibitors of human beta-secretase 1, and HIV for predicting inhibitor activity. We trained random forest models with a 5-fold cross-validation strategy to ensure reproducible results. The tokenized form of the molecules was converted into one-hot encoding, which was used as the input feature representation for training the models. The regression tasks were evaluated using the root mean squared error (RMSE) metric, while receiver operating characteristic area under the curve (ROC-AUC) was used for evaluating the classification tasks.

Our findings indicate that the choice of tokenization scheme has a significant impact on the performance of molecular property prediction models. In terms of regression tasks, the AIS tokenization scheme demonstrated superior performance, achieving the lowest RMSE values on ESOL (0.553), FreeSolv (0.441), and Lipophilicity (0.683) datasets. In terms of classification tasks, AIS performed strongly, with the highest ROC-AUC values achieved on BBBP (0.885), and the second-highest value achieved on BACE (0.835) and HIV (0.729) datasets. In Table 5, the performances of SMILES, DeepSMILES, and SELFIES schemes were found to be comparable across all the conducted tasks, with SELFIES exhibiting a slightly more favorable performance than SMILES. However, SmilesPE exhibited inconsistent results compared to other tokenization schemes. While it performed the best on the HIV and BACE datasets and performed well on the BBBP and Lipophilicity datasets, its performance was not consistent across all datasets (worst in ESOL and FreeSolv). This inconsistency could be attributed to the limited vocabulary size of SmilesPE, resulting in an inaccurate representation of molecular substructures. Moreover, the representation of molecules as sequences of pseudo-words in SmilesPE may not be suitable for certain types of molecules, leading to poorer performance on specific datasets.

AIS as a fingerprint

Due to the compositional nature of the AIS tokenization scheme, it can be readily converted to a fingerprint by simply removing non-atomic tokens. We considered the frequency of each AIS token from a string as its fingerprint form. This is conceptually identical to counting the frequency of each distinct fragment during ECFP

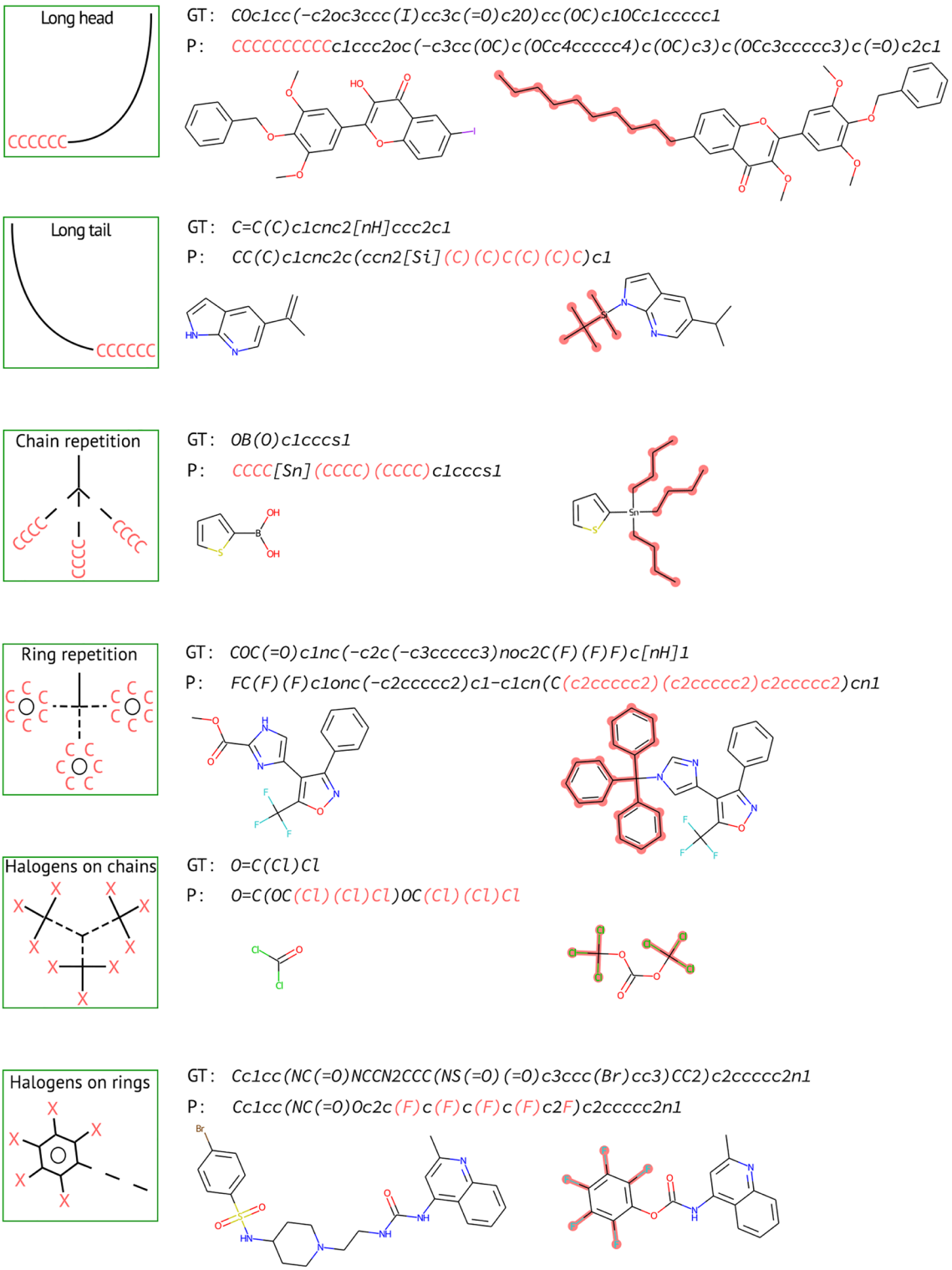

Fig. 6 The most commonly occurring repetitive patterns in an NMT retrosynthetic framework. Copious repetitions are highlighted in SMILES and molecular drawings. GT and P refer to the ground truth and prediction, respectively

Table 4 Performance (top-1 accuracy) of various tokenization schemes on single-step retrosynthesis task and the number of predictions with token repetition

Tokenization schemes	rep-l$\|_P$ − rep-l$\|_{GT}$ ≥ 2	Acc.(%) greedy	
		String exact	Tc exact
Atom-wise baseline [57]	–	42.00	–
Atom-wise (ref. [57] is reproduced)	801	42.05	44.72
SmilesPE (ref. [21])	821	19.82	22.74
SELFIES (ref. [17])	886	28.82	30.76
DeepSMILES (ref. [16])	902	38.63	41.20
Atom-in-SMILES	727	46.32	47.62

Table 5 Performance analysis of tokenization schemes for molecular property prediction using MoleculeNet benchmark suite

	SMILES	DeepSMILES	SELFIES	SmilesPE	AIS
Regression Datasets: **RMSE**					
ESOL	0.628	0.631	0.675	0.689	**0.553**
FreeSolv	0.545	0.544	0.564	0.761	**0.441**
Lip	0.924	0.895	0.938	0.800	**0.683**
Classification Datasets: **ROC-AUC**					
BBBP	0.758	0.777	0.799	0.847	**0.885**
BACE	0.740	0.774	0.746	**0.837**	0.835
HIV	0.649	0.648	0.653	**0.739**	0.729

Comparison of Random Forest regression and classification models with 5-Fold Cross-Validation. Bold emphasis denotes the highest performing approach

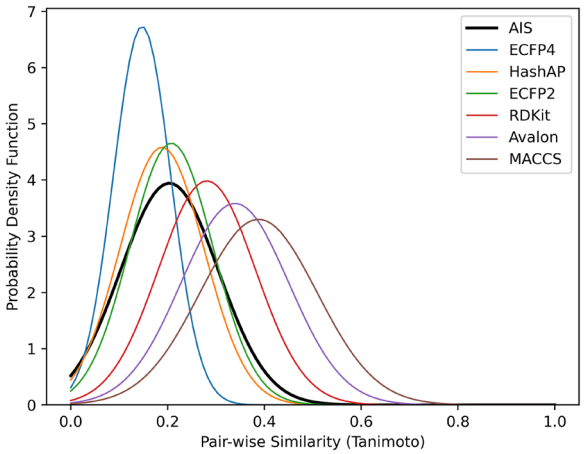

Fig. 7 Fingerprint nature of AIS. Pairwise similarity scores of 1 million pairs of molecules are computed for the commonly used structural fingerprints and their probability density functions are plotted. The Tanimoto coefficient is used to measure similarity scores

generation [50]. It should be noted that computing the molecular similarity with AIS does not require any hashing function. Thus, converting an AIS string to its fingerprint form requires much less computation than other fingerprint methods using hash functions. Based on this definition, we calculated the Tanimoto similarities of 2 million pairs generated by pairwise combination of 2000 randomly chosen ChEMBL molecules using the AIS fingerprint and other widely used fingerprint schemes, and their probability densities are compared (Fig. 7). The most probable similarity between a random pair of molecules using the AIS fingerprint is 0.21. This is similar to those of HashAP and ECFP2 and lower than those of RDKit, Avalon, and MACCS. This indicates that the AIS fingerprint has better resolution power than MACCS, Avalon, and RDKit, and comparable to ECFP2 and HashAP.

This similarity between AIS and its fingerprint form may enhance the learning process of various chemical language models. In general, the loss functions of chemical translation and generation models are assessed through a token-wise comparison. However, few errors in a SMILES string may lead to an invalid or substantially different molecule. Consequently, the loss value and molecular similarities may not be closely correlated. On the contrary, AIS strings with a few token errors represent similar molecules because of the fingerprint-like nature of AIS. Thus, loss values and dissimilarities of molecules due to token errors are more closely correlated with AIS than SMILES.

In a recent study, we established that fingerprint representations, such as ECFP2, ECFP4, and atom environments, can be transformed back into their corresponding SMILES strings with minimal ambiguity [44]. This suggests that fingerprint representations can serve as valuable and informative stand-alone representations. Employing fingerprints as input representations simplifies the application of diverse AI models to chemistry, as bit vectors or straightforward token counts are more manageable than character sequences and can be effortlessly integrated with numerous existing algorithms. We contend that the strong resemblance between AIS strings and their fingerprint counterparts holds significant potential for further development in this domain.

Conclusion

This study demonstrated that tokenization has a significant impact on the final prediction quality. We introduced atom-in-SMILES (AIS) tokenization as a proper and meaningful custom tokenization scheme to improve the prediction quality in sequence prediction tasks achieving gains of up to 10.7% in equivalent SMILES mapping and 4.3% in a retrosynthetic prediction task.

AIS outperforms other tokenization methods in molecular property prediction tasks and aligns more closely with chemical perspectives.

We investigated the resolution of the fingerprint aspect of AIS, revealing that it encompasses all essential information for seamless bidirectional transitions between SMILES and fingerprint representations, ensuring practical implementation. The study addressed the repetition issue in molecular predictions, akin to natural language, which impeded the quality of predicted molecules. The AIS tokenization scheme considerably diminished obstacles in repetitive loops (by around 10%) in the predicted SMILES. As far as we are aware, no prior research has examined token degeneration in AI-driven chemical applications. The AIS tokenization method can be employed by the broader community to deliver chemically precise and customized tokens for molecular prediction, property prediction, and generative models.

Author contributions

UVU, IA and JL conceived and designed the study. UVU and IA processed data, trained the models and analyzed the results. UVU, IA., and JL discussed and interpreted the results, wrote and reviewed the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (Nos. NRF-2019M3E5D4066898, NRF-2020M3A9G7103933, and NRF-2022R1C1C1005080). This research was supported by the BK21FOUR Program of the National Research Foundation of Korea(NRF) funded by the Ministry of Education(5120200513755). This work was also supported by the Korea Environment Industry & Technology Institute (KEITI) through the Technology Development Project for Safety Management of Household Chemical Products, funded by the Korea Ministry of Environment (MOE) (KEITI:2020002960002 and NTIS:1485017120).

Availability of data and materials

The source code of this work is available via GitHub repo: https://github.com/snu-lcbc/atom-in-SMILES

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 16 January 2023 Accepted: 14 May 2023
Published online: 29 May 2023

References

1. 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
2. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comp Sci 28(1):31–36. https://doi.org/10.1021/ci00057a005
3. Bader RFW (1985) Atoms in molecules. Acc Chem Res 18(1):9–15. https://doi.org/10.1021/ar00109a003
4. Cadeddu A, Wylie EK, Jurczak J, Wampler-Doty M, Grzybowski BA (2014) Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. Angew Chem Int Ed 53(31):8108–8112. https://doi.org/10.1002/anie.201403708
5. Lesniewski S (1927) O podstawach matematyki (on the foundations of mathematics). Przeglad filozoficzny 30:164–206
6. Varzi AC (1996) Parts, wholes, and part-whole relations: the prospects of mereotopology. Data Knowl Eng 20(3):259–286. https://doi.org/10.1016/S0169-023X(96)00017-1
7. Borbély G, Kornai A (2019) Sentence Length. arXiv Preprint. https://doi.org/10.48550/arXiv.1905.09139
8. Bojar O, Chatterjee R, Federmann C, Graham Y, Haddow B, Huck M, Jimeno Yepes A, Koehn P, Logacheva V, Monz C, Negri M, Névéol A, Neves M, Popel M, Post M, Rubino R, Scarton C, Specia L, Turchi M, Verspoor K, Zampieri M (2016) Findings of the 2016 conference on machine translation. In: proceedings of the first conference on machine translation: volume 2, shared task papers, pp. 131–198. Association for Computational Linguistics, Berlin, Germany . https://doi.org/10.18653/v1/W16-2301
9. Holtzman A, Buys J, Du L, Forbes M, Choi Y (2019) The curious case of neural text degeneration. arXiv Preprint. https://doi.org/10.48550/arXiv.1904.09751
10. Welleck S, Kulikov I, Roller S, Dinan E, Cho K, Weston J (2019) Neural text generation with unlikelihood training. arXiv Preprint. https://doi.org/10.48550/arXiv.1908.04319
11. Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond JL, Chen H, Engkvist O (2019) Randomized SMILES strings improve the quality of molecular generative models. J Cheminform 11(1):1–13. https://doi.org/10.1186/s13321-019-0393-0
12. Lin T-S, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z, Woods E, Craig SL, Johnson JA, Kalow JA, Jensen KF, Olsen BD (2019) Bigsmiles: a structurally-based line notation for describing macromolecules. ACS Cent Sci 5(9):1523–1531. https://doi.org/10.1021/acscentsci.9b00476
13. Drefahl A (2011) CurlySMILES: a chemical language to customize and annotate encodings of molecular and nanodevice structures. J Cheminform 3(1):1–7. https://doi.org/10.1186/1758-2946-3-1
14. ChemAxon Extended SMILES and SMARTS - CXSMILES and CXSMARTS - Documentation. https://docs.chemaxon.com/display/docs/chemaxon-smiles-extensions.md. Accessed: 10 Feb 2022
15. OpenSMILES. Home page http://opensmiles.org. Accessed: 10 Dec 2021
16. O'Boyle NM, Dalke A (2018) DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. ChemRxiv. https://doi.org/10.26434/chemrxiv.7097960.v1
17. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A (2020) Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach Learn Sci Technol 1(4):045024. https://doi.org/10.1088/2632-2153/aba947
18. O'Boyle NM (2012) Towards a universal SMILES representation - a standard method to generate canonical SMILES based on the InChI. J Cheminform 4(9):1–14. https://doi.org/10.1186/1758-2946-4-22
19. Schneider N, Sayle RA, Landrum GA (2015) Get your atoms in order-an open-source implementation of a novel and robust molecular canonicalization algorithm. J Chem Inf Model 55(10):2111–2120. https://doi.org/10.1021/acs.jcim.5b00543
20. Hähnke VD, Bolton EE, Bryant SH (2015) PubChem atom environments. J Cheminform 7(1):1–37. https://doi.org/10.1186/s13321-015-0076-4
21. Li X, Fourches D (2021) SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning. J Chem Inf Model 61(4):1560–1569. https://doi.org/10.1021/acs.jcim.0c01127
22. 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
23. 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/
24. Lample G, Conneau A (2019) Cross-lingual language model pretraining. arXiv. https://doi.org/10.48550/arXiv.1901.07291
25. Quirós M, Gražulis S, Girdzijauskaitė S, Merkys A, Vaitkus A (2018) Using SMILES strings for the description of chemical connectivity in the crystallography open database. J Cheminform 10(1):23. https://doi.org/10.1186/s13321-018-0279-6
26. Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N, Müller K (2009) Benchmark data set for in silico prediction of

ames mutagenicity. J Chem Inform Model. https://doi.org/10.1021/ci900161g
27. O'Donnell VB, Dennis EA, Wakelam MJO, Subramaniam S (2019) LIPID MAPS: serving the next generation of lipid researchers with tools, resources, data, and training. Sci Signal 12(563):2964. https://doi.org/10.1126/scisignal.aaw2964
28. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X (2013) Use of natural products as chemical library for drug discovery and network pharmacology. PLoS ONE 8(4):62839. https://doi.org/10.1371/journal.pone.0062839
29. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. Nucleic Acids Res 39(suppl–1):1035–1041. https://doi.org/10.1093/nar/gkq1126
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv. Neural Inf. Process Syst. 2017–Decem:5999–6009
31. Bahdanau D, Cho KH, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 1–15
32. Blum LC, Reymond J-L (2009) 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc 131(25):8732–8733. https://doi.org/10.1021/ja902302h
33. Blum LC, Deursen Rv, Reymond J-L (2011) Visualisation and subsets of the chemical universe database GDB-13 for virtual screening. J Comput Aided Mol Des 25(7):637–647. https://doi.org/10.1007/s10822-011-9436-y
34. GDB-13 Database. Home page https://gdb.unibe.ch/downloads/. Accessed: 02 Nov 2022
35. Ucak UV, Ji H, Singh Y, Jung Y (2016) A soft damping function for dispersion corrections with less overfitting. J. Chem. Phys. 145(17):174104. https://doi.org/10.1063/1.4965818
36. Segler MHS, Waller MP (2017) Neural-symbolic machine learning for retrosynthesis and reaction prediction. Eur J Chem 23(25):5966–5971. https://doi.org/10.1002/chem.201605499
37. Jin W, Coley CW, Barzilay R, Jaakkola T (2017) Predicting organic reaction outcomes with weisfeiler-lehman network. Adv Neural Inf Process Syst 2017–Decem:2608–2617
38. Coley CW, Green WH, Jensen KF (2018) Machine learning in computer-aided synthesis planning. Acc Chem Res 51(5):1281–1289. https://doi.org/10.1021/acs.accounts.8b00087
39. 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 3(10):1103–1113. https://doi.org/10.1021/acscentsci.7b00303
40. Karpov P, Godin G, Tetko IV (2019) A transformer model for retrosynthesis. In: artificial neural networks and machine learning – ICANN 2019: workshop and special sessions, pp. 817–830. Springer, Cham
41. Tetko IV, Karpov P, Van Deursen R, Godin G (2020) State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. Nat Commun 11(1):1–11. https://doi.org/10.1038/s41467-020-19266-y
42. Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, Bekas C, Iuliano A, Laino T (2020) Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chem Sci 11(12):3316–3325. https://doi.org/10.1039/c9sc05704h
43. Ucak UV, Kang T, Ko J, Lee J (2021) Substructure-based neural machine translation for retrosynthetic prediction. J Cheminform 13(1):1–15. https://doi.org/10.1186/s13321-020-00482-z
44. Ucak UV, Ashyrmamatov I, Ko J, Lee J (2022) Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. Nat Commun 13(1):1186. https://doi.org/10.1038/s41467-022-28857-w
45. Szymkuć S, Gajewska EP, Klucznik T, Molga K, Dittwald P, Startek M, Bajczyk M, Grzybowski BA (2016) Computer-assisted synthetic planning: the end of the beginning. Angew Chem Int Ed 55(20):5904–5937. https://doi.org/10.1002/anie.201506101
46. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF (2017) Prediction of organic reaction outcomes using machine learning. ACS Cent Sci 3(5):434–443. https://doi.org/10.1021/acscentsci.7b00064
47. Law J, Zsoldos Z, Simon A, Reid D, Liu Y, Khew SY, Johnson AP, Major S, Wade RA, Ando HY (2009) Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. J Chem Inf Model 49(3):593–602. https://doi.org/10.1021/ci800228y
48. Lowe DM (2012) Extraction of chemical structures and reactions from the literature. PhD thesis, University of Cambridge. https://doi.org/10.17863/CAM.16293
49. Lowe D (2017) Chemical reactions from US patents (1976-Sep2016). Figshare. https://doi.org/10.6084/m9.figshare.5104873.v1
50. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754. https://doi.org/10.1021/ci100050t
51. Rajan K, Steinbeck C, Zielesny A (2022) Performance of chemical structure string representations for chemical image recognition using transformers. Digit Discov 1(2):84–90. https://doi.org/10.1039/d1dd00013f
52. Nair P, Singh AK (2021) On reducing repetition in abstractive summarization. In: proceedings of the student research workshop associated with RANLP 2021, pp. 126–134. INCOMA Ltd., Online. Accessed 17 Apr 2023 https://aclanthology.org/2021.ranlp-srw.18
53. Jawahar G, Abdul-Mageed M, Lakshmanan LVS (2020) Automatic detection of machine generated text: A critical survey. In: proceedings of the 28th international conference on computational linguistics, pp. 2296–2309. International Committee on Computational Linguistics, Barcelona, Spain (Online). Accessed 17 Apr 2023 https://doi.org/10.18653/v1/2020.coling-main.208.https://aclanthology.org/2020.coling-main.208
54. Ferruz N, Schmidt S, Höcker B (2022) A deep unsupervised language model for protein design. BioRxiv. https://doi.org/10.1101/2022.03.09.483666
55. Edwards C, Lai T, Ros K, Honke G, Cho K, Ji H (2022) Translation between molecules and natural language. arXiv. https://doi.org/10.48550/arxiv.2204.11817
56. Fu Z, Lam W, So AM-C, Shi B (2020) A theoretical analysis of the repetition problem in text generation. arXiv. https://doi.org/10.48550/arxiv.2012.14660
57. Lin K, Xu Y, Pei J, Lai L (2020) Automatic retrosynthetic route planning using template-free models. Chem Sci 11(12):3355–3364. https://doi.org/10.1039/c9sc03666k
58. Wu Z, Ramsundar B, Feinberg E, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018) Moleculenet: a benchmark for molecular machine learning. Chem Sci 9:513–530. https://doi.org/10.1039/C7SC02664A

Publisher's Note