scientific reports

Check for updates

OPEN

Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling

Miguelangel Leon[1,3], Yuriy Perezhohin[1,3], Fernando Peres[1], Aleš Popovič[2] & Mauro Castelli[1]✉

Life sciences research and experimentation are resource-intensive, requiring extensive trials and considerable time. Often, experiments do not achieve their intended objectives, but progress is made through trial and error, eventually leading to breakthroughs. Machine learning is transforming this traditional approach, providing methods to expedite processes and accelerate discoveries. Deep Learning is becoming increasingly prominent in chemistry, with Convolutional Graph Networks (CGN) being a key focus, though other approaches also show significant potential. This research explores the application of Natural Language Processing (NLP) to evaluate the effectiveness of chemical language representations, specifically SMILES and SELFIES, using tokenization methods such as Byte Pair Encoding (BPE) and a novel approach developed in this study, Atom Pair Encoding (APE), in BERT-based models. The primary objective is to assess how these tokenization techniques influence the performance of chemical language models in biophysics and physiology classification tasks. The findings reveal that APE, particularly when used with SMILES representations, significantly outperforms BPE by preserving the integrity and contextual relationships among chemical elements, thereby enhancing classification accuracy. Performance was evaluated in downstream classification tasks using three distinct datasets for HIV, toxicology, and blood–brain barrier penetration, with ROC-AUC serving as the evaluation metric. This study highlights the critical role of tokenization in processing chemical language and suggests that refining these techniques could lead to significant advancements in drug discovery and material science.

The convergence of computational chemistry and data science has transformed how chemical structures are represented and analyzed. In this evolving field, accurately and efficiently representing chemical information is critical for advancing drug discovery, material science, and other life sciences. Traditional machine learning models, while effective, often struggle with the complexity of chemical structures, highlighting the need for methods that can better decompose and interpret chemical languages. Recent advancements in Natural Language Processing (NLP) and BERT-based transformer models[1] present a promising opportunity to address these challenges by adapting tokenization techniques from natural language to chemical language.

This work is motivated by the growing need for more precise tokenization strategies that preserve the integrity of chemical structures[2]. Existing tokenization methods, such as Byte Pair Encoding (BPE)[3], have limitations[4] when applied to chemical languages like Simplified Molecular Input Line Entry System (SMILES)[5] and SELF Referencing Embedded Strings (SELFIES)[6]. These techniques often fail to capture the contextual relationships necessary for accurate molecular representation. To address this gap, we propose a novel tokenization approach called Atom Pair Encoding (APE), specifically designed to improve the classification of chemical structures in BERT-based models. The primary contributions of this study are as follows:

- We introduce the APE tokenizer, a method tailored for chemical languages, and demonstrate its superiority over traditional BPE in maintaining the structural integrity of chemical representations.
- We compare the performance of BPE and APE tokenization methods using SMILES and SELFIES representations in transformer-based models, showing that APE significantly improves the accuracy of molecular classification tasks.
- Our evaluation, based on biophysics and physiology datasets, reveals that models utilizing APE tokenization outperform state-of-the-art approaches, providing a new benchmark for chemical language modeling.These findings suggest that refining tokenization techniques can lead to significant advancements in computational chemistry[7–10], opening up new possibilities for accelerating research in drug discovery, material science, and

[1]NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, 1070-312 Lisbon, Portugal. [2]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia. [3]These authors contributed equally: Miguelangel Leon and Yuriy Perezhohin. ✉email: mcastelli@novaims.unl.pt

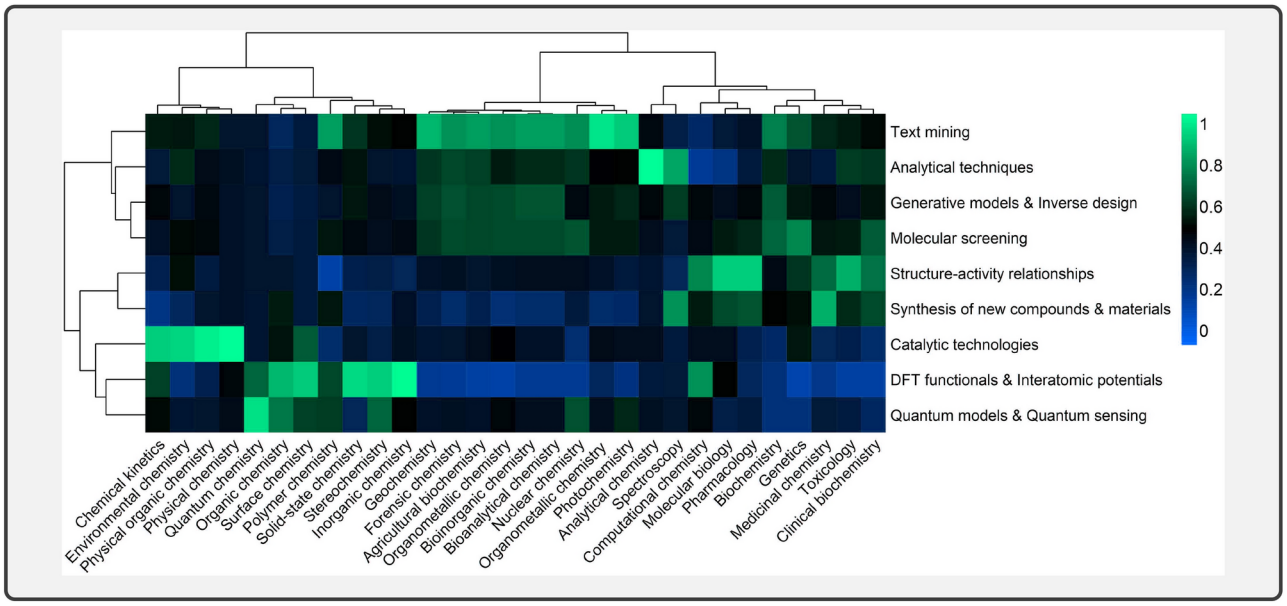

Fig. 1. A comprehensive overview of the impact of machine learning in various branches of chemistry. The heatmap visualizes the frequency of machine learning results in these areas, based on data from 2008 to June 30, 2019. Green (1) is the most significant contribution and blue (0) is the least significant[10].

Name/link	Description
Text2Concrete	Developing Sustainable Concretes with In-Context Learning
Molecule discovery by context	ScholarBERT to find potential hydrogen carrier molecules by analyzing their scientific context.
Extracting structured data from free-form organic synthesis text	Extracting Structured Data from Organic Synthesis Procedures Using a Fine-Tuned Large Language Model
sMolTalk	A proof-of-concept that uses OpenAI's GPT-3.5 to generate 3dmol.js code based on natural language input.

Table 1. An overview of some NLP models applied to chemistry, developed in a hackathon by Jablonka et al.[16].

other branches of machine learning tasks. Figure 1 illustrates the growing influence of machine learning across various branches of chemistry, further motivating the need for advanced tokenization strategies to enhance model performance in chemical informatics tasks.

Related work

Natural language processing and chemistry

Before the advent of Large Language Models (LLM), Jastrzebski et al.[11] made a groundbreaking contribution by establishing a link between NLP and cheminformatics. They demonstrated the potential of NLP techniques, particularly using recurrent neural networks[12] (RNN), to solve classification problems. Their study focused on predicting activity against target proteins, a crucial step in computer-aided drug design. Utilizing representations like SMILES, their research highlighted the effectiveness of NLP methods and paved the way for future studies in this area.

Jiang et al.[13] advanced cheminformatics by employing machine learning techniques to decode and predict chemical phenomena. They introduced an approach that utilized SMILES to analyze text data encoded as chemical reactions in SMILES format. This approach demonstrated the flexibility and effectiveness of long short-term memory[14] (LSTM) networks in handling chemical data and predicting the practicality and yield of organic synthesis reactions. The model's success in managing a range of reactions demonstrates the potential of combining NLP with deep learning, creating new possibilities for automating and streamlining drug discovery and chemical synthesis processes. Additionally, this study emphasized the need for dynamic and adaptable techniques to fully capture the details of chemical interactions, setting a standard for future innovations in this field.

Noike et al.[15] and Jablonka[16] made significant strides by incorporating LLMs into the field of chemistry. They employed these models to comprehend chemical literature, predict reaction outcomes, and even propose hypotheses for experimentation (Table 1). Such automation has the potential to expedite discoveries in chemistry by processing large amounts of data and intricate patterns that may be challenging for humans to handle efficiently. These advancements highlight the versatility and potential of LLMs in driving research, particularly in materials science and chemistry domains[17,18].

More recently, Jablonka et al.[19] explored the use of large language models (LLMs), specifically GPT-3[20], in an open question-answer format for chemical queries. The authors fine-tuned GPT-3 across diverse tasks such as classification, regression, and inverse design. In classification and regression, the model was tasked with providing answers in the form of specific labels or numerical values, respectively. In inverse design, the model had to generate representations of specific molecules or reactions based on given prompts. The authors concluded that the fine-tuning process yielded positive results, even outperforming conventional machine learning techniques in scenarios with limited data, thereby demonstrating the ability of LLMs to transfer knowledge to chemical tasks. White et al.[21] examined LLMs trained on code-generation tasks, showing that these models can effectively interpret chemistry-related problems. The research focused on posing chemistry problems as coding tasks and evaluated them based on the correctness of the code and the evaluation of field experts. The study concluded that LLMs can understand a variety of chemistry tasks when posed as coding problems, and it is possible to augment their accuracy by 30% through specific prompt engineering strategies. Furthermore, the authors have described best practices for using LLMs in chemical tasks and open-sourced their dataset and evaluation tools.

While LLMs have shown strong performance across domains, some researchers[22,23] argue that limitations in accessing external knowledge can hinder their effectiveness in scientific research. Bran et al.[24] addressed this concern by introducing ChemCrow, a framework that enhances GPT-4[25] with 18 integrated tools, such as Web Search, Name2Smiles, SynthesisPlanner, and SynthesisExecuter. This complex structure improves performance on chemistry tasks and demonstrates the framework's potential to automate a wide range of scientific activities.

All in all, the combination of LLMs and chemistry is transforming how scientific research is conducted, analyzed, and produced[26,27]. The adaptability of these models in tasks like classification, regression, and generation, along with principles from NLP, creates new opportunities for innovation and exploration in materials science and chemistry[2].

Text representation of chemical compounds

Chemical structures can be represented in various ways, including sequences of atoms and graph structures (see Fig. 2). In machine learning, diverse representations are employed to capture the intricate chemical properties of molecules. These representations underlie tasks in cheminformatics, drug discovery, and quantitative structure-activity relationship (QSAR) studies.

SMILES, a string-based notation introduced by Weininger[5], is one of the most used methods for representing chemical structures.

SMILES offers a concise and human-readable format for representing chemical structures using text strings. By leveraging ASCII characters to depict atoms and bonds within a molecule, SMILES facilitates the exchange and analysis of chemical information by researchers. This simplicity has led to its widespread adoption in cheminformatics databases, such as PubChem[28].

However, despite its extensive use, SMILES notation does have limitations, as discussed by Krenn et al.[6]:

- *Robustness in Generative Models:* SMILES can generate semantically invalid strings when used in generative models. This often results in several invalid molecule outputs, hampering automated approaches to molecule design and discovery.
- *Inconsistency with Isomers:* Another challenge associated with SMILES is to consistently represent isomers. The SMILES notation can lead to a single SMILES string corresponding to multiple molecules, or conversely, different strings potentially representing the same molecule. This ambiguity can hinder database searches and introduce complications in comparative studies.
- *Difficulty with Certain Chemical Classes:* SMILES sometimes struggle to represent certain chemical classes like organometallic compounds or complex biological molecules.To address the limitations of SMILES in cheminformatics and machine learning applications, SELFIES was developed as a string-based representation for graphs. Unlike SMILES, every SELFIES string guarantees a molecule representation without semantic errors. This robustness is crucial in computational chemistry applications in molecule design using models like Variational Auto-Encoding (VAE)[6].

Experiments have shown that SELFIES consistently produces molecules with random mutations of valid strings, like MDMA (commonly known as ecstasy). In contrast, SMILES often generates invalid strings when mutated[6].

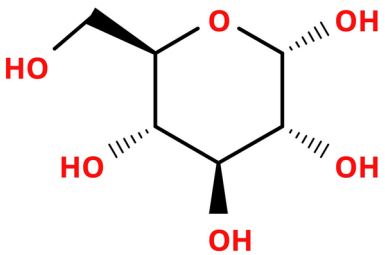

Fig. 2. D-Glucose: A 3D representation of the sugar molecule, where wedge bonds indicate bonds projecting towards the viewer, and hashed bonds indicate bonds extending away from the viewer. SMILES notation: C(C1C(C(C(C(O1)O)O)O)O)O, SELFIES notation: [C][Branch2][Ring1][Branch1][C][C][Branch1][S][C][Branch1][N][C][Branch1][Branch2][C][Branch1][Ring2][O][Ring1][=Branch1][O][O][O][O][O].

Model	BBBP ROC	Tox21 ROC
D-MPNN	0.697	0.719
RF	0.7194	0.724
GCN	0.676	0.688
ChemBERTa-1	0.733	
ChemBERTa-2		
MLM-5M	0.701	0.762
MLM-10M	0.696	0.748
MLM-77M	0.698	0.749
MTR-5M	0.742	0.834
MTR-10M	0.733	0.827
MTR-77M	0.728	0.817

Table 2. Comparative results of different deep learning models as presented by Ahmad at al.[36] on selected MoleculeNet datasets[29].

Aspects	Natural language	SMILES/SELFIES
Sequence length	15–20 words	45–60
Token space	> 100k	1000 smaller
Token order	Tone, meaning, fluency	Different molecules

Table 3. Comparison between natural and chemical languages[42]. Although chemical sequences tend to be longer, the vocabulary is smaller, and tokens occur with greater frequency.

The use of SELFIES in models such as VAE has proven advantageous. It enables these models to convert graphs into continuous representations, which can then be optimized using gradient-based or Bayesian methods. Models based on SELFIES have shown enhancements in the diversity and complexity of molecules. The latent space of SELFIES-based VAE is denser than that of SMILES by two orders of magnitude. This implies a more comprehensive exploration of the chemical space during optimization procedures.

Overall, SELFIES represents a significant advancement in cheminformatics. It offers a robust, flexible, and comprehensive solution for representing structures in computational chemistry applications, particularly those involving machine learning and AI-driven molecular design.

Transformers-based models

Despite the success of Convolutional Graph Networks (CGNs) in predicting molecular properties[29–31], data collection remains the main obstacle because experimental data generation is expensive and time-consuming.

To overcome this issue, one can leverage unlabeled data for knowledge extraction. Transformers[32] have emerged as a powerful tool for self-supervised learning from text. Their development has been accelerated by libraries like Hugging Face[33] and advancements in powerful GPUs for tensor calculations. With vast databases containing millions of molecule strings[28], transformers present an opportunity for processing massive data, particularly using Masked Language Modeling (MLM) for pre-training, common in BERT architectures[1].

Several pre-trained transformer models for molecules have been developed[34]. These models utilized SMILES strings, MLM, and BPE for tokenization. Fine-tuned with MoleculeNet datasets[29], they achieved comparable, but not superior, results to baseline models from Chemprop[35]. Subsequent iterations with larger datasets, like the 77 million SMILES used in ChemBERTa's second iteration[36], yielded no significant improvements. Even models using SELFIES for pre-training achieved similar results[37]. This consistency across models (see Table 2) with the same sub-word tokenization (BPE) suggests the need for further exploration.

Tokenization of molecules' strings

Molecule strings like SMILES and SELFIES are designed for computer interpretation, not human readability. While smaller molecules are interpretable with SMILES, complex structures become unwieldy. However, this benefit for computers presents a challenge for applying NLP techniques, which often rely on tokenization for numerical representation.

A popular NLP method, Byte Pair Encoding (BPE)[34,36–38], was originally introduced for data compression[3] and later adapted for NLP[39]. BPE excels at breaking down words into meaningful units, handling large vocabularies, and managing uncommon words - all crucial for NLP tasks.

However, applying BPE to chemical languages like SMILES and SELFIES presents unique challenges. These languages have significantly smaller vocabularies (roughly 1000 times smaller) compared to natural languages (see Table 3) and represent chemical structures. While tokenizing each character might seem straightforward, limitations arise. Chemical languages like SMILES and SELFIES use a limited character set repeatedly, with characters representing structures and functional groups depending on context and position. This complexity

makes capturing the role of atoms (identified by atomic number) in different positions difficult. Simply splitting chemical strings into individual characters may oversimplify the structures[40].

Recognizing and addressing this challenge is crucial when applying NLP techniques to chemical languages. Order and structure within SMILES and SELFIES necessitate a tokenization approach that captures the relationships and roles of characters in defining how atoms function within molecules. While BPE has been successful in NLP, directly applying it to chemical languages might not fully capture the complexity of structures. This is because BPE simplifies representations by breaking down sequences without considering the impact of structural context and character position in chemical strings[41].

This challenge intensifies when training deep learning models with chemical languages. Models may struggle to learn and predict chemical properties or reactions if the tokenization method oversimplifies structures and loses crucial information about the molecule. Therefore, while BPE and its adaptations show promise in NLP, their application in chemical language processing requires careful consideration and potential adjustments.

A data-driven substructural tokenization algorithm called SMILES Pair Encoding (SPE) has been proposed for deep learning applications in chemistry[42]. Inspired by the BPE method, SPE iteratively merges substrings within SMILES strings to create tokens. This approach allows SPE to recognize and represent chemical substructures as distinct and meaningful units.

SPE offers two key benefits:

1. *Focus on Chemical Substructures*: By extracting and representing common SMILES substrings as unique tokens, SPE emphasizes meaningful chemical building blocks. These tokens capture more comprehensive information about molecular functionality compared to individual atom-level tokens.
2. *Compact Representation for Deep Learning*: SPE generates compact input sequences for deep learning models. This reduces computational demands and accelerates training speed (see Fig. 3).Thus, the advantages of SPE are twofold. First, it captures the richness of chemical information by encoding meaningful substructures within unique tokens. These tokens, representing recurring SMILES substrings, provide a more comprehensive and nuanced representation of molecular functions compared to individual atom-based tokens. Second, SPE offers a streamlined representation for deep learning models, reducing the computational burden and speeding up the training process, as illustrated in Fig. 3.

The principles behind SPE can be extended to SELFIES strings as well. The inherent robustness of SELFIES, where every string guarantees a valid molecule, offers the potential for 100% valid substructures within tokens - a limitation of SMILES.

In SELFIES, the fundamental unit is represented by a bracketed string, such as [C], [Ring1], or [=O]. This characteristic makes SELFIES a promising candidate for applying SPE's substructure-focused tokenization. This adaptation could further enhance and refine how molecular data is represented and analyzed within machine learning frameworks.

Methodology

This research investigates and compares tokenization methods for chemical string representations, SMILES and SELFIES. The core focus lies in evaluating the impact of a novel tokenizer on these different representations in downstream NLP tasks.

The paper introduces an Atom Pair Encoding (APE) tokenizer, drawing inspiration from both BPE and SPE. The underlying hypothesis is that a tokenizer specifically designed for chemical representations will yield better results when applied to SMILES and SELFIES data in downstream NLP tasks.

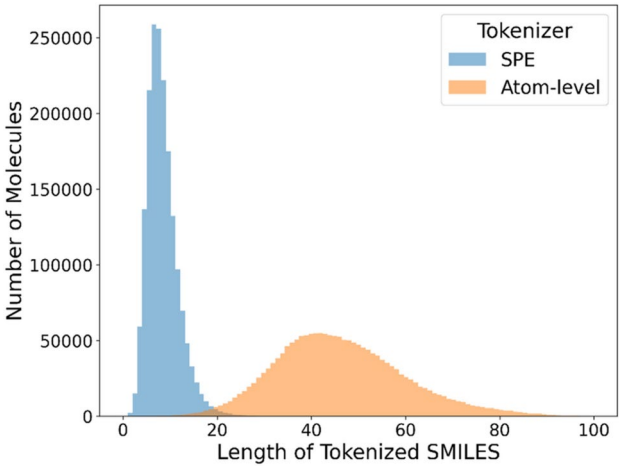

Fig. 3. Length distribution of the tokenized SMILES using CheMBL dataset. Orange represents atom-level tokenization and blue SPE tokenization. Reprinted with permission from *J.Chem. Inf. Model.*[42] Copyright 2021, American Chemical Society.

The methodology starts with the tokenization of chemical strings, where SMILES and SELFIES representations are processed using both the traditional BPE and the proposed APE. This step examines how each tokenization method decomposes chemical strings into subword units, focusing on their ability to preserve the critical structural information necessary for accurate chemical modeling.

Next, the pre-training phase utilizes these tokenized representations across four models based on the RoBERTa[43] architecture. Specifically, the models-SMILES-BPE, SMILES-APE, SELFIES-BPE, and SELFIES-APE-are pre-trained using their respective tokenizers to learn from the chemical language data. The objective of this step is to prepare these models for fine-tuning by ensuring that they can effectively process the tokenized chemical strings.

In the fine-tuning phase, the models are fine-tuned and evaluated on three specific downstream tasks: Blood–Brain Barrier Penetration (BBBP)[44], Drug Therapeutics Program AIDS Antiviral Screen (HIV)[45], and Toxicology of the 21st Century (Tox21)[46]. These tasks are designed to assess the models' ability to classify chemical structures based on their biological activity and physiological properties.

Finally, a comparative analysis is performed to evaluate the performance of each model across the downstream tasks. This step compares the strengths and weaknesses of the BPE and APE tokenizers, offering insights into which method better retains molecular structure and improves classification accuracy.

A visual representation of the methodology flow is provided in Fig. 4.

Tokenization

Two tokenization strategies were applied to both SMILES and SELFIES:

1. *BPE*: Standard subword tokenizer used as a base tokenizer.
2. *APE*: A tokenizer tailored for SMILES and SELFIES.The APE tokenizer breaks down SMILES and SELFIES strings into fundamental units that capture chemical meaning. For example, the string "[C][=C][C][=C][C][=C][Ring1][=Branch1] would be split into [[C], [=C], [C], [=C], [C], [=C], [Ring1], [=Branch1]]" as basic units before starting merging the tokens iteratively, similar to the BPE tokenizer, given as result two tokens "[C][=C][C][=C] and [C][=C][Ring1][=Branch1]". The resulting tokens depend on the vocabulary used to train the tokenizer.

Given the internal limitations of the SPE tokenizer, it was not included in the evaluation. In response to these limitations, APE was developed to overcome three significant drawbacks of SPE:

- Despite claims from the authors, the SPE Python library does not work with SELFIES.
- In its current state, the SPE library produces a maximum of approximately 3000 tokens, whereas APE, using the same data, generates around 5300 different tokens. Both tokenizers were used with a minimum frequency of 2000 for merging pairs, a shared hyperparameter.A public dataset from PubChem[28] containing 10 million SMILES strings was used for training the tokenizers. SELFIES strings were generated for this dataset using the selfies library[6]. This library transforms SMILES strings into an intermediary graph representation before converting them to SELFIES. Each tokenizer was trained on a subset of 2 million molecules from the PubChem dataset. The BPE tokenizers leveraged the Hugging Face library[33] for training.

This results in four distinct tokenizers:

- SMILES-BPE
- SMILES-APE
- SELFIES-BPE
- SELFIES-APE

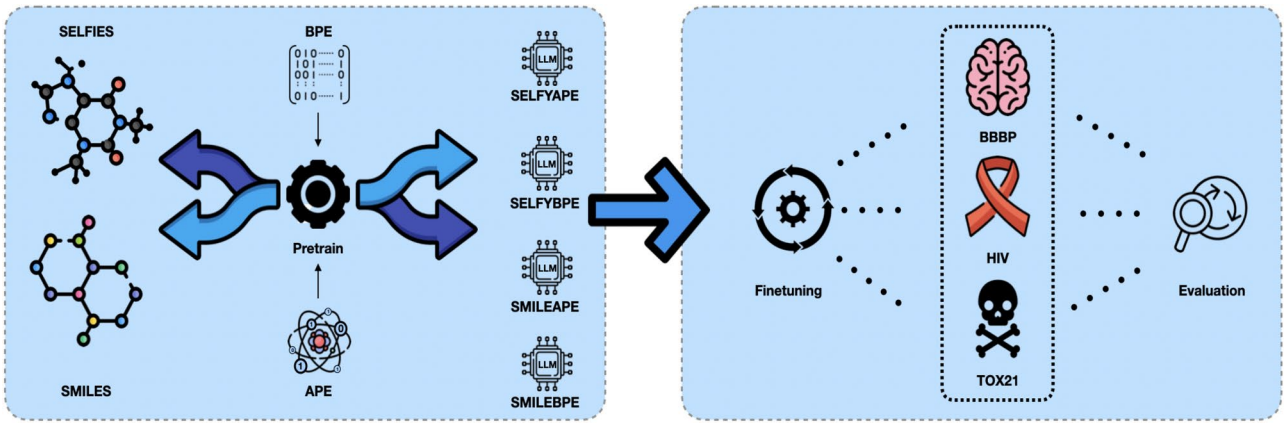

Fig. 4. Overview of the general methodology flow.

Category	Dataset	Data type	Task type	# Tasks	# Compounds	Metric
Biophysics	HIV	SMILES	Classification	1	41,127	ROC-AUC
Physiology	BBBP	SMILES	Classification	1	2039	ROC-AUC
Physiology	Tox21	SMILES	Classification	12	7831	ROC-AUC

Table 4. MoleculeNet datasets details.

	Model	BBBP ROC	HIV ROC	Tox21 ROC
Text based	SMILYAPE-1M	**0.754 ± 0.006**	0.772 ± 0.010	0.838 ± 0.002
	SMILYBPE-1M	0.746 ± 0.006	0.754 ± 0.015	0.849 ± 0.002
	SELFYAPE-1M	0.735 ± 0.015	0.768 ± 0.012	0.842 ± 0.002
	SELFYBPE-1M	0.676 ± 0.014	0.709 ± 0.012	0.825 ± 0.001
	ChemBERTa-2-MTR-77M	0.698 ± 0.014	0.735 ± 0.008	0.790 ± 0.003
	SELFormer	0.716 ± 0.021	0.769 ± 0.010	0.838 ± 0.005
Graph based	MoleculeNet-Graph-Conv	0.690	0.763	0.829
	D-MPNN	0.737	**0.776**	**0.851**

Table 5. Results and comparisons with different graph-based and text-based models on MoleculeNet benchmarks. APE models achieve comparable scores and, in most cases, outperform D-MPNN and MoleculeNet-Graph-Conv. Bold used to identify the best values.

Model pre-training

In this phase, each tokenizer was used to pre-train a corresponding RoBERTa-based model. The models all shared a common architecture with 6 hidden layers, a hidden size of 768, an intermediate size of 1536, and 12 attention heads. A batch size of 32 was used, with 15% of the tokens masked. To optimize hyperparameters, a search was conducted using the Optuna library[47].

The training dataset comprised 1 million molecules from the PubChem dataset, with a validation set of 100,000 molecules. Each model was pre-trained for 20 epochs using the AdamW optimizer[48]. The computations were performed on an NVIDIA 3060 GPU with 12GiB of VRAM.

Fine-tuning and evaluation

The performance of the models was evaluated on three curated datasets from MoleculeNet[29]. These datasets were chosen for their relevance to drug discovery tasks:

- Blood–Brain Barrier Penetration[44]
- Drug Therapeutics Program AIDS Antiviral Screen[45]
- Toxicology of the 21st Century[46]All datasets originally contained SMILES strings for molecule representation. Similar to the pre-training dataset, these strings were converted to SELFIES using the selfies library. Each dataset was then divided into training, validation, and testing sets using an 80/10/10 split, following MoleculeNet's recommendations. The DeepChem library[49] was used to handle data loading and splitting.

A summary of the datasets can be found in Table 4.

The evaluation metric used was the mean area under the curve (ROC-AUC) of the Receiver Operating Characteristic (ROC) curve. This metric aligns with MoleculeNet's recommendations, as it is commonly used for evaluating models on these datasets and facilitates direct comparisons. All models underwent fine-tuning for five epochs, with early stopping based on the ROC-AUC score on the validation set. The final evaluation was performed on the held-out test set.

Results

The ROC-AUC scores for the three datasets are presented in Table 5, with visualizations in Fig. 5. These scores are compared against the performance of two text-based transformer models: ChemBERTa (second iteration)[36], which uses SMILES strings, and SELFormer[37], which utilizes SELFIES strings. Additionally, two graph-based models serve as baselines: chemprop's D-MPNN[50] and MoleculeNet Graph-Conv[29]. The ChemBERTa and SELFormer models were obtained from the Hugging Face Hub and evaluated locally ten times for consistency. The results for the graph-based models are those reported by their respective authors.

The results show that models tokenized with APE generally outperform those using BPE. Moreover, models utilizing SMILES strings often achieved better performance than those using SELFIES strings. Notably, the SMILYAPE model outperformed the state-of-the-art chemprop model on the BBBP dataset and exhibited competitive performance on the other datasets, as shown in Table 5.

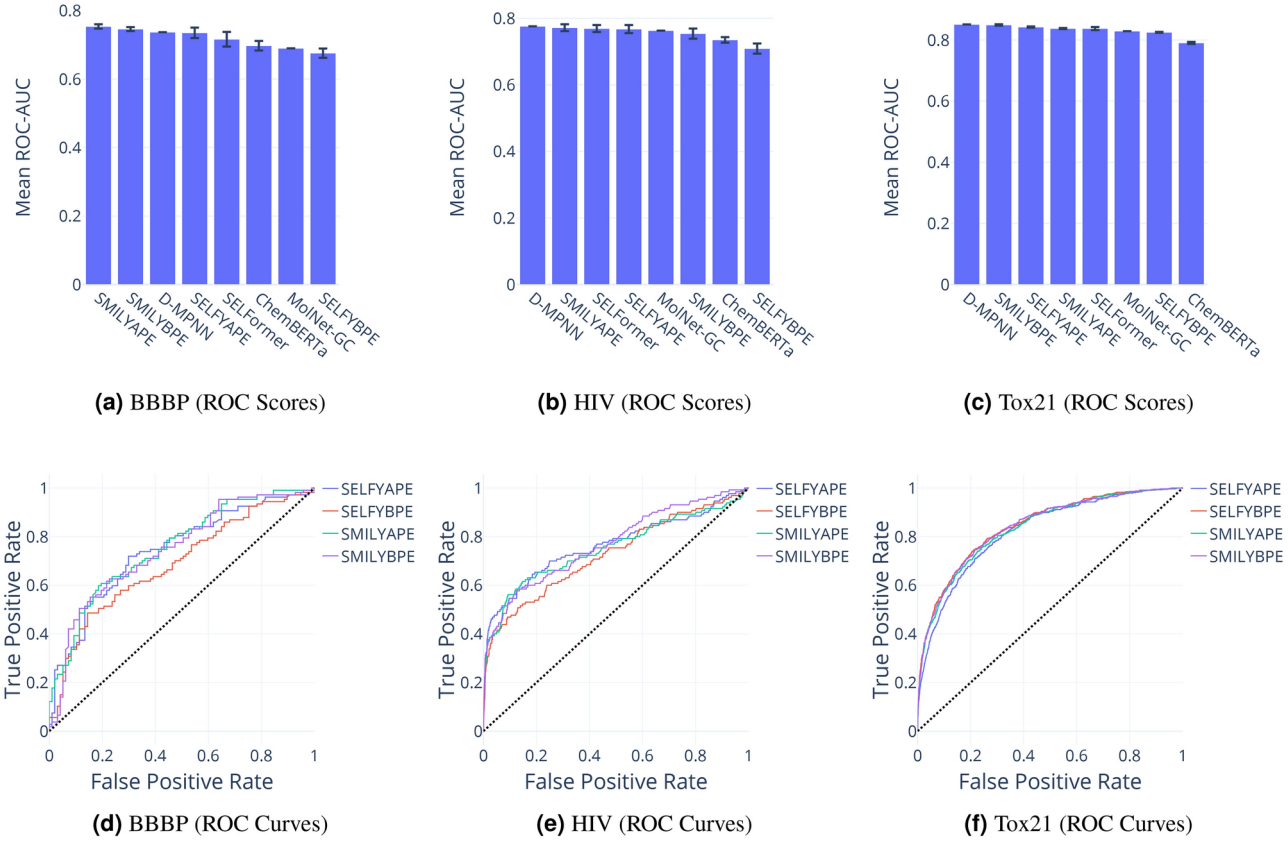

Fig. 5. ROC-AUC scores and curves of selected benchmarks. Top row: (**a**) BBBP, (**b**) HIV, and (**c**) Tox21 bar charts. Bottom row: (**d**) BBBP, (**e**) HIV, and (**f**) Tox21. The Tox21 is the largest dataset, giving better results with all models.

	BBBP	HIV	Tox21
SMILES Δ	0.008	0.018	− 0.012
SELFIES Δ	0.059	0.059	0.017

Table 6. Difference in ROC-AUC scores on selected datasets between SMILYAPE and SMILYBPE, as well as between SELFYAPE and SELFYBPE. The SELFIES-based models show greater variations in scores.

Model	Accuracy	Precision	Recall	f1 score
SELFYAPE	0.6520	0.6875	0.6168	0.6502
SELFYBPE	0.6716	0.7564	0.5514	0.6378
SMILYAPE	0.6961	0.7922	0.5701	0.6630
SMILYBPE	0.6863	0.8209	0.5140	0.6322

Table 7. Accuracy, precision, recall, and f1 score for a balanced dataset, BBBP.

Atom pair encoding versus byte pair encoding

The results consistently favor models tokenized with APE over those using BPE, likely due to BPE's tendency to perform character-level pairing. This can result in issues such as:

- *Stray characters*: Pairings like "C)(" might occur, creating meaningless tokens.
- *Splitting elements*: Chemical elements such as Chlorine ("Cl") might be split into "C" and "l", misrepresenting it as carbon and an unrelated "l," which could then be incorrectly paired with other characters.In contrast, APE prioritizes preserving the identity of atoms, ensuring a more accurate representation of chemical context within the tokens. This advantage is particularly pronounced for SELFIES-based models, as each fundamen-

Model	Accuracy	Precision	Recall	f1 score
SELFYAPE	0.9691	0.5231	0.2615	0.3487
SELFYBPE	0.9682	0.4906	0.2000	0.2842
SMILYAPE	0.9691	0.5205	0.2923	0.3744
SMILYBPE	0.9677	0.4789	0.2615	0.3383

Table 8. Accuracy, precision, recall, and f1 score for a imbalanced dataset, HIV with 3.74% of the true class.

Dataset	U value	*p* value	Cliff's delta
BBBP	13	0.006	0.74
HIV	15	0.009	0.70
Tox21	0	0.002	− 1.00

Table 9. Statistical analysis of SMILYAPE and SMILYBPE on selected benchmarks.

tal unit is enclosed in brackets, preventing the splitting issues seen with BPE. This likely explains the greater performance difference observed in models using SELFIES strings (see Table 6 for details).

To compare the performance of the models across both the balanced (BBBP) and imbalanced (HIV) datasets, it is important to focus on several metrics-accuracy, precision, recall, and F1 score-considering the context of the dataset's balance or imbalance.

Balanced dataset BBBP (Table 7)

- *Accuracy*: The best performing model based on accuracy is SMILYAPE (69.61%), followed by SELFYBPE (67.16%). This suggests that APE-based models perform better than BPE-based models in terms of overall classification performance.
- *Precision*: In terms of precision, SMILYBPE achieves the highest score (82.09%), indicating it is the most effective at avoiding false positives.
- *Recall*: SELFYAPE achieves the highest recall (61.68%), indicating it is the best at identifying true positives.
- *F1 Score*: Regarding the F1 score, SMILYAPE performs the best (66.30%), followed by SELFYAPE (65.02%).The best overall performing model, considering all metrics, is SMILYAPE. It has the highest accuracy and F1 score, making it a balanced choice that handles both precision and recall well. Although SMILYBPE has the highest precision, its low recall pulls down its overall performance.

Imbalanced Dataset HIV (3.74% Positive Class, Table 8)
For the imbalanced HIV dataset, accuracy is less informative because the majority class dominates the dataset. In this scenario, precision, recall, and F1 score are more meaningful metrics since they give insight into how well the model handles the minority class.

- *Accuracy*: All models have very similar accuracy (around 96.7–96.9%), reflecting the influence of the majority class.
- *Precision*: SELFYAPE (52.31%) and SMILYAPE (52.05%) have the highest precision, indicating that they make fewer false positive errors when predicting the minority class.
- *Recall*: SMILYAPE also has the highest recall (29.23%), indicating it is the best at identifying positive cases. However, all models have relatively low recall, showing that they struggle to identify the minority class, which is typical in imbalanced datasets.
- SMILYAPE achieves the highest F1 score (37.44%), followed closely by SELFYAPE (34.87%). This suggests that SMILYAPE is the most effective at handling the imbalanced nature of the dataset, as it strikes a better balance between precision and recall compared to the other models.The best-performing model on the imbalanced HIV dataset is SMILYAPE, as it has the highest recall, precision, and F1 score. Although SELFYAPE has slightly better precision, SMILYAPE outperforms it overall due to its superior recall and F1 score.

To statistically assess the observed performance differences between SMILYAPE and SMILYBPE models (Tables 5 and 6), non-parametric tests were employed due to the limited number of evaluation runs (10) which likely resulted in a non-normal distribution of scores. The Mann-Whitney U test[51] and Cliff's delta[52] were chosen for this purpose, as they do not rely on the assumption of normality. The detailed results are presented in Table 9.

The consistently low p-values (below 0.05) across all datasets indicate a statistically significant difference in performance between the SMILYAPE and SMILYBPE models. This suggests that the APE tokenizer offers a meaningful advantage over BPE for tasks involving chemical string representations.

The values of Cliff's delta deviate significantly from zero, indicating a substantial difference in performance between the tokenization methods. In the Tox21 dataset, the negative delta (− 1.00) suggests that SMILYBPE outperforms SMILYAPE. However, for the BBBP and HIV datasets, the positive delta values (0.74 and 0.70, respectively) indicate that SMILYAPE performs better than SMILYBPE.

Furthermore, the analysis for SELFYAPE compared to SELFYBPE consistently shows a lack of overlap between the two samples. This is evident from the consistently positive Cliff's delta values for all three datasets (1.00 for BBBP, 1.00 for HIV, and 1.00 for Tox21), indicating that SELFYAPE achieves statistically superior results.

SMILES versus SELFIES

Both SMILYAPE and SELFYAPE models prioritize local context, with greater emphasis on immediate neighbors within the chemical representation (Fig. 6). However, visual inspection of a small sample of molecules reveals some degree of attention to distant tokens for both models. An analysis of attention weights revealed that SMILYAPE, on average, allocated more attention to both immediate neighboring tokens and self-attention than SELFYAPE. Specifically, SMILYAPE assigned a weight of 0.108 to immediate neighbors compared to 0.096 for SELFYAPE. Conversely, SMILYAPE exhibited lower attention towards distant tokens (0.030) than SELFYAPE (0.043). This finding aligns well with the superior performance of SMILYAPE, as chemical bonding is primarily determined by the directly connected atoms, with a diminishing effect from more distant ones.

The inherent differences between SMILES and SELFIES representations lead to variations in sequence length for each molecule. Tokenization further contributes to this disparity: SMILYAPE consistently generates a lower average number of tokens than SELFYAPE for the same molecule (8.6 tokens vs. 11.9 tokens, respectively). This efficiency in tokenization potentially allows the SMILYAPE model to allocate attention more effectively. By focusing on a smaller number of potentially more informative tokens, SMILYAPE might achieve enhanced performance in downstream tasks.

A deeper analysis of the predictions on the BBBP dataset reveals that all true positives identified by SELFYBPE were also correctly predicted by SELFYAPE, with SELFYAPE achieving a higher recall than SELFYBPE (61.68% vs. 55.14%). The overlap in true positive predictions between the two models suggests that SELFIES-based models tend to capture similar molecular information despite using different tokenization strategies. This finding supports the hypothesis that SELFIES encoding offers a robust and consistent representation of molecular structures, allowing both APE and BPE tokenization methods to identify the same true positives. The fact that SELFYAPE, using APE, captures all the true positives identified by SELFYBPE suggests that, while BPE is more flexible in compressing tokens, it does not lose significant structural information captured by APE. This overlap likely stems from SELFIES' inherent ability to generalize molecular features, allowing both tokenizers to extract similar patterns across diverse molecules. Ultimately, the observation that different tokenization approaches can lead to similar true positive predictions reinforces the idea that SELFIES-based models are highly effective in capturing core structural information across diverse molecules, regardless of the tokenizer used.

An analysis of the true positives overlap between the SMILES-based models revealed that they share only 29.304% of the true positives. This overlap quantifies the proportion of actual positives correctly identified by both models, indicating that the APE and BPE tokenization strategies capture distinct yet partially overlapping

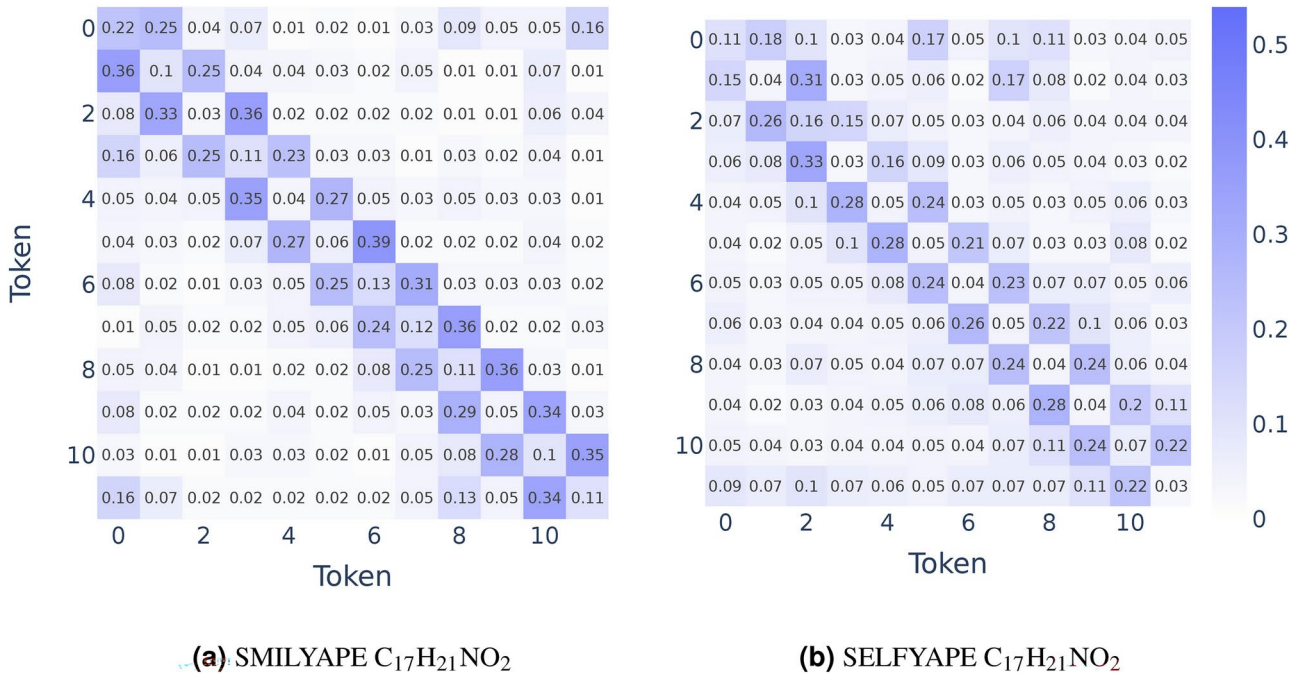

Fig. 6. Attention patterns of the SMILYAPE (**a**) and SELFYAPE (**b**) models, respectively, for molecule $C_{17}H_{21}NO_2$ from the BBBP dataset. The visualization highlights a stronger focus on attention between adjacent tokens (darker squares along the diagonal). However, some attention is also directed towards distant tokens (lighter squares), suggesting the models capture both local context and broader semantic relationships within the molecule.

molecular information. This relatively low overlap in true positives suggests that SMILES-based models are more sensitive to tokenization differences than SELFIES-based models, where the overlap in true positive predictions was more pronounced. The 29.304% intersection indicates that while both SMILYAPE and SMILYBPE effectively identify positive instances, they do so in a complementary manner, with each model capturing unique patterns that the other may overlook. This highlights the significant influence of tokenization strategies on model predictions, particularly in models using traditional SMILES representations.

SELFIES encoding ensures that molecular substructures are represented in a way that directly corresponds to the chemical structure, avoiding the syntactic ambiguity often present in SMILES. SMILES-based models, however, tend to be more sensitive to the tokenization method used. Since SMILES strings rely on a precise sequence of atoms and bonds, different tokenization techniques can lead to significant variations in how molecules are parsed and represented. As a result, SMILES-based models may show less overlap in true positive predictions, with each model capturing different aspects of a molecule based on how the SMILES string is tokenized. Despite this sensitivity often being viewed as a limitation, it can also be a strength. SMILES-based datasets can be augmented to enhance diversity and data size, allowing for more robust model training. In this study, SMILES-based models generally outperformed their SELFIES-based counterparts, highlighting the potential advantages of exploiting SMILES' flexibility in molecular representation.

Conclusions

This study investigated chemical language modeling through a comparative analysis of two key representations: SMILES, the traditional standard, and the emerging SELFIES. We employed RoBERTa-based transformer models and evaluated the effectiveness of two tokenization strategies: Byte Pair Encoding (BPE) and Atom Pair Encoding (APE). Our methodological approach and analysis revealed that APE consistently outperforms BPE, particularly in enhancing the processing and modeling of chemical languages using SMILES representations.

The detailed investigation highlighted a key advantage of APE: its ability to preserve the structural integrity and connections between elements within molecules. This leads to improvements in model performance across various chemical informatics classification tasks. Notably, models trained and fine-tuned with APE tokenization consistently outperformed those using BPE, for both SMILES and SELFIES representations. These findings highlight the critical role of tokenization strategies in chemical language processing, where the choice of representation and tokenization can significantly influence computational model outcomes.

A promising direction for future research lies in refining tokenization methods to capture chemical functional groups more accurately. By putting greater emphasis on identifying and representing these functional groups, future tokenization techniques could achieve higher levels of detail and precision in chemical language modeling. This advancement could lead to models that not only excel in classification tasks but also demonstrate proficiency in predicting molecular properties, reactions, and interactions with greater accuracy.

In conclusion, this research advances the the understanding of chemical language modeling by examining the effectiveness of SMILES and SELFIES representations in conjunction with tokenization techniques. We demonstrated the advantages of APE for enhancing RoBERTa-based models in chemical informatics and initiated a broader discussion on the crucial role of tokenization and representation in computational chemistry. Further exploration of these concepts offers significant potential for advancements in drug discovery, material science, and other scientific fields, highlighting the transformative potential of computational methods on natural sciences research.

Data availability

The datasets generated and/or analysed during the current study are available in the Hugging Face repository https://huggingface.co/datasets/mikemayuare/PubChem10M_SMILES_SELFIES

Received: 21 June 2024; Accepted: 14 October 2024
Published online: 23 October 2024

References

1. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). ArXiv: 1810.04805.
2. Guo, T. et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. *Adv. Neural Inf. Process. Syst.* **36**, 59662–59688 (2023).
3. Gage, P. A new algorithm for data compression. *C Users J.* **12**, 23–38. https://doi.org/10.5555/177910.177914 (1994).
4. Tran, K. Optimization of molecular transformers: Influence of tokenization schemes. M.Sc. Thesis, Chalmers University of Technology, 2021 (2021).
5. Weininger, D. SMILES, a chemical language and information system: 1: Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36. https://doi.org/10.1021/ci00057a005 (1988).
6. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **1**, 045024. https://doi.org/10.1088/2632-2153/aba947 (2020).
7. Mater, A. C. & Coote, M. L. Deep learning in chemistry. *J. Chem. Inf. Model.* **59**, 2545–2559. https://doi.org/10.1021/acs.jcim.9b00266 (2019).
8. Goh, G. B., Hodas, N. O. & Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **38**, 1291–1307. https://doi.org/10.1002/jcc.24764 (2017).
9. Jiao, Z., Hu, P., Xu, H. & Wang, Q. Machine learning and deep learning in chemical health and safety: A systematic review of techniques and applications. *ACS Chem. Health Saf.* **27**, 316–334. https://doi.org/10.1021/acs.chas.0c00075 (2020).
10. Cova, T. F. G. G. & Pais, A. A. C. C. Deep learning for deep chemistry: Optimizing the prediction of chemical patterns. *Front. Chem.* **7**, 809. https://doi.org/10.3389/fchem.2019.00809 (2019).
11. Jastrzębski, S., Leśniak, D. & Czarnecki, W. M. Learning to SMILE(S), arXiv:1602.06289 (2016). _eprint: 1602.06289.
12. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).

13. Jiang, S. et al. When SMILES smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access* **9**, 85071–85083. https://doi.org/10.1109/ACCESS.2021.3083838 (2021).
14. Hochreiter, S. *Long Short-Term Memory* (Neural Computation MIT-Press, 1997).
15. Boiko, D. A., MacKnight, R. & Gomes, G. Emergent autonomous scientific research capabilities of large language models. ArXiv:abs/2304.05332 (2023).
16. Jablonka, K. M. et al. 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. *Digit. Discov.* **2**, 1233–1250. https://doi.org/10.1039/D3DD00113J (2023).
17. Xia, J., Zhu, Y., Du, Y. & Li, S. Z. A systematic survey of chemical pre-trained models. arXiv preprint (2022). arXiv:2210.16484.
18. Liao, C., Yu, Y., Mei, Y. & Wei, Y. From words to molecules: A survey of large language models in chemistry. arXiv preprint arXiv:2402.01439 (2024).
19. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
20. Brown, T. B. Language models are few-shot learners. arXiv preprint (2020). arXiv:2005.14165.
21. White, A. D. et al. Assessment of chemistry knowledge in large language models that generate code. *Digit. Discov.* **2**, 368–376 (2023).
22. Schick, T. *et al.* Toolformer: Language models can teach themselves to use tools. *Adv. Neural Inf. Process. Syst.* **36** (2024).
23. Shuster, K. *et al.* Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. arXiv preprint (2022). arXiv:2203.13224.
24. Bran, M. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
25. Achiam, J. *et al.* Gpt-4 technical report. arXiv preprint (2023). arXiv:2303.08774.
26. Castro Nascimento, C. M. & Pimentel, A. S. Do large language models understand chemistry? A conversation with Chatgpt. *J. Chem. Inf. Model.* **63**, 1649–1655 (2023).
27. White, A. D. The future of chemistry is language. *Nat. Rev. Chem.* **7**, 457–458 (2023).
28. Kim, S. et al. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395. https://doi.org/10.1093/nar/gkaa971 (2021).
29. Wu, Z. et al. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530. https://doi.org/10.1039/c7sc02664a (2018).
30. Daller, E., Bougleux, S., Brun, L. & Lézoray, O. Local patterns and supergraph for chemical graph classification with convolutional networks. In *Structural, Syntactic, and Statistical Pattern Recognition* (eds Bai, X. et al.) 97–106 (Springer International Publishing, 2018).
31. Ryu, S., Lim, J., Hong, S. H. & Kim, W. Y. Deeply learning molecular structure-property relationships using attention- and gate-augmented graph convolutional network. arXiv: Learning (2018).
32. Vaswani, A. *et al.* Attention is all you need (2023). arXiv:1706.03762.
33. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) 38–45, https://doi.org/10.18653/v1/2020.emnlp-demos.6 (Association for Computational Linguistics, Online, 2020).
34. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. https://doi.org/10.09885arXiv:Learning (2020).
35. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.*[SPACE]https://doi.org/10.1021/ACS.JCIM.9B00237 (2019).
36. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta-2: Towards chemical foundation models (2022). arXiv:2209.01712.
37. Yüksel, A., Ulusoy, E., Ünlü, A. & Doğan, T. SELFormer: Molecular representation learning via SELFIES language models. *Mach. Learn.: Sci. Technol.* **4**, 025035. https://doi.org/10.1088/2632-2153/acdb30 (2023).
38. Cao, Z. et al. MOFormer: Self-supervised transformer model for metal-organic framework property prediction. *J. Am. Chem. Soc.*[SPACE]https://doi.org/10.1021/JACS.2C11420 (2023).
39. Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Erk, K. & Smith, N. A.) 1715–1725, https://doi.org/10.18653/v1/P16-1162 (Association for Computational Linguistics, Berlin, Germany, 2016).
40. Bader, R. F. W. Atoms in molecules. *Acc. Chem. Res.* **18**, 9–15. https://doi.org/10.1021/ar00109a003 (1985).
41. Ucak, U. V., Ashyrmamatov, I. & Lee, J. 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 (2023).
42. Li, X. & Fourches, D. SMILES pair encoding: A data-driven substructure tokenization algorithm for deep learning. *J. Chem. Inf. Model.* **61**, 1560–1569. https://doi.org/10.1021/acs.jcim.0c01127 (2021).
43. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach (2019). arXiv:1907.11692.
44. Martins, I. F., Teixeira, A. L., Pinheiro, L. & Falcão, A. O. A Bayesian approach to in silico blood–brain barrier penetration modeling. *J. Chem. Inf. Model.* **52**, 1686–1697. https://doi.org/10.1021/ci300124c (2012).
45. National Cancer Institute. AIDS Antiviral Screen Data (2024).
46. National Institutes of Health. Tox21 Challenge (2014).
47. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019).
48. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. (2019) arXiv:1711.05101.
49. Ramsundar, B. et al. *Deep Learning for the Life Sciences* (O'Reilly Media, 2019).
50. Heid, E. et al. Chemprop: A machine learning package for chemical property prediction. *J. Chem. Inf. Model.* **64**, 9–17. https://doi.org/10.1021/acs.jcim.3c01250 (2024).
51. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60. https://doi.org/10.1214/aoms/1177730491 (1947).
52. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.* **114**, 494–509. https://doi.org/10.1037/0033-2909.114.3.494 (1993).

Acknowledgements

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS). Aleš Popovič was supported by the Slovenian Research and Innovation Agency (ARIS) under research core funding P2-0442.

Author contributions

All authors designed the study. M.L. developed the algorithms and wrote the software. M.L. performed the experiments. Y.P. and F.P. validated the results. M.C. supervised the project with input from F.P. and A.P. All the

authors drafted the paper. All authors reviewed the final paper.

Competing interests
The authors declare no competing interests.

Models and Algorithm
The APE tokenizer is hosted on GitHub https://github.com/mikemayuare/apetokenizerThe pretrained models and the fined-tuned ones can be found on Hugging Face https://huggingface.co/mikemayuare

Additional information
Correspondence and requests for materials should be addressed to M.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.