Heliyon 10 (2024) e39038

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

Transformer-based models for chemical SMILES representation: A comprehensive literature review

Medard Edmund Mswahili, Young-Seob Jeong *

Chungbuk National University, Department of Computer Engineering, Cheongju, 28644, South Korea

ARTICLE INFO

Keywords:
Transformers
Natural language processing
SMILES
Machine learning
Cheminformatics
BERT
GPT

ABSTRACT

Pre-trained chemical language models (CLMs) have attracted increasing attention within the domains of cheminformatics and bioinformatics, inspired by their remarkable success in the natural language processing (NLP) domain such as speech recognition, text analysis, translation, and other objectives associated with language. Furthermore, the vast amount of unlabeled data associated with chemical compounds or molecules has emerged as a crucial research focus, prompting the need for CLMs with reasoning capabilities over such data. Molecular graphs and molecular descriptors are the predominant approaches to representing molecules for property prediction in machine learning (ML). However, Transformer-based LMs have recently emerged as de-facto powerful tools in deep learning (DL), showcasing outstanding performance across various NLP downstream tasks, particularly in text analysis. Within the realm of pre-trained transformer-based LMs such as, BERT (and its variants) and GPT (and its variants) have been extensively explored in the chemical informatics domain. Various learning tasks in cheminformatics such as the text analysis that necessitate handling of chemical SMILES data which contains intricate relations among elements or atoms, have become increasingly prevalent. Whether the objective is predicting molecular reactions or molecular property prediction, there is a growing demand for LMs capable of learning molecular contextual information within SMILES sequences or strings from text inputs (i.e., SMILES). This review provides an overview of the current state-of-the-art of chemical language Transformer-based LMs in chemical informatics for de novo design, and analyses current limitations, challenges, and advantages. Finally, a perspective on future opportunities is provided in this evolving field.

1. Introduction

SMILES [1], which stands for simplified molecular input line entry system, is a line notation that allows a user to represent and describe a chemical structure or a chain of letters, numbers and characters that specify the atoms, their connectivity, bond order and chirality in a way that can be used by the computer for various tasks such as natural language processing (NLP) related tasks [2]. SMILES strings are composed of ASCII characters and provide a textual representation of molecular structures. Consider the example of SMILES notation for Paracetamol is CC(=O)Nc1ccc(O)cc1. In SMILES representation atoms are represented by the their atomic symbols. Metal atoms are represented with the symbols in square brackets ('[]'), however, brackets can be omitted for B, C, N, O, P, S, and Halides. Aromatic C, O, S, and N atoms are shown in lower case 'c', 'o', 's', and 'n'. Aliphatic C, O, S, and N atoms are

* Corresponding author.
E-mail address: ysjay@chungbuk.ac.kr (Y.-S. Jeong).

https://doi.org/10.1016/j.heliyon.2024.e39038
Received 13 June 2024; Received in revised form 26 September 2024; Accepted 5 October 2024

Available online 9 October 2024

shown in upper case 'C', 'O', 'S', and 'N'. On the other hand single, double, and triple bonds are represented by the symbols (-, =, and #) respectively. Adjacent atoms are assumed to be connected to each other by a single or aromatic bond. Branches are specified by enclosing them in parentheses, which can be nested or arranged. SMILES is widely used to represent chemical structures in text form, efficiently encoding molecular information and facilitating chemical data interoperability [3].

Driven by novel reactions and diverse building blocks, the fully enumerated REAL database has grown significantly from approximately 170 million compounds in 2017 to over 5.5 billion compounds in 2022 [4]. It now forms the majority of the popular and widely-used ZINC20 virtual screening database [5]. Chemical structure representation such as SMILES, is a critical aspect of computational chemistry or cheminformatics, aiding in the analysis and prediction of several tasks such as molecular properties, compound-protein interaction, and de novo molecular generation in the field of drug discovery. Recently, deep learning (DL) techniques have been widely adopted to accelerate and improve drug discovery and have delivered impressive results in various applications, such as molecular property prediction, drug–drug interaction (DDI) prediction, drug–target interaction (DTI) prediction, drug repositioning, and de novo drug generation [6–9]. However, the problem of limited labeled data problem is still challenging and also restricts the performance of these techniques in specific tasks. Additionally, data labeling is a resource-intensive and time-consuming task to be accomplished in every stage of drug discovery pipeline [10].

Based on the probably approximately correct learning theory, language models (LMs) heavily depend on the size of their training data or in other words, larger training datasets assist the model in making highly accurate predictions. Therefore, an effective and promising approach involves leveraging both a vast number of unlabeled data or molecules and the limited available labeled data, wherein CLMs acquire insights or more informative molecular representation through pre-training (i.e., unsupervised pre-training) with unlabeled data (molecules) and fine-tuning with labeled data for specific downstream tasks. LMs leverage and customize algorithms originally designed for natural language processing (NLP) to understand the language of chemistry. This adaptation is facilitated by the use of string representations, such as SMILES strings. [3].

Transformer-based models have revolutionized numerous fields within artificial intelligence (AI), most notably in NLP, where they have demonstrated unparalleled capabilities in understanding and generating human language [11]. This same architecture, known for its ability to capture long-range dependencies and contextual information, has recently been adapted for use in computational chemistry focusing on their use with SMILES representations to enhance the prediction of molecular properties [12,13]. Here, it facilitates the interpretation and prediction of chemical properties from SMILES strings, a text-based representation of chemical structures [14]. Suitability of transformer-based models can be seen in terms of the criteria such as: Diverse Applications: Transformer-based models are well-suited for a wide range of molecular properties analysis tasks, including drug discovery, toxicity prediction, and materials science. State-of-the-Art Performance: With appropriate fine-tuning and optimization, transformer models have demonstrated state-of-the-art performance across various molecular property prediction benchmarks. Flexibility: Transformers can accommodate different types of molecular representations beyond SMILES, such as graph-based representations, enabling their application to diverse cheminformatics tasks. Continual Improvement: As transformer architectures evolve and new training methodologies emerge, their suitability for molecular properties analysis is expected to improve further, addressing current limitations and expanding their applicability in the field.

Over the years, various methods have been employed to encode molecular structures, with recent advancements focusing on transformer and attention-based models. Consequently, this literature review aims to provide a comprehensive overview of the application of transformer and attention-based models in computational chemistry handling SMILES representations to facilitate the prediction of molecular properties. In this study, we aim to provide a thorough review of various transformer (i.e., attention) based CLMs as well as a systematic taxonomy of the applications in handling SMILES representations. To summarize, our contributions are:

- We provide a detailed review over existing transformer based CLMs. We present a general design pipeline and discuss the variants of each module based on transformer architecture.
- We discuss major applications and their corresponding methods contributed to various chemical properties predictions.
- We propose open problems for future research and provide a thorough analysis of each problem and propose future research directions.

2. Background

This section discusses the limitations of traditional molecular representation approaches and highlights the need for more advanced models to capture intricate relationships within molecular structures.

Before the breakthrough of transformer and attention-based models, traditional molecular representation methods such as molecular descriptors, fingerprints, and graph-based representations were predominant approaches to representing molecules. These are also common methods previously employed for predicting molecular properties which often utilize machine learning (ML) models. While effective, these methods often required expert-engineered or pre-designed molecular features representation and extensive domain knowledge to enhance predictive performance [15]. Moreover, numerous research studies have incorporated explicit 3D atomic coordinates to achieve even higher performance levels [15]. Another significant research direction focuses on optimizing the model architecture, whether it operates on molecular descriptors or fingerprints, directly processes SMILES strings, or analyzes the underlying molecular graph [16].

The effectiveness of DL methods relies greatly on having large-scale labeled training data. While various research areas such as image processing often reaches millions of labeled samples, molecular property prediction faces a severely different scenario. Gathering such a vast number of labeled molecular properties through screening experiments is exceptionally costly and labor-intensive. This resembles the challenge in natural language modeling, where there's an abundance of unlabeled data but only a tiny is labeled. To address this, the prevailing approach is the pre-training and fine-tuning framework or semi-supervised learning architecture [17]. This method involves initially training the model in an unsupervised manner, then fine-tuning it with labeled data. Consequently, few unsupervised fingerprint methods have been presented previously as a solution to the challenge of limited labeled data.

The seq2seq et al. [18] and Seq3seq et al. [19] fingerprint models are the pioneered semi-supervised frameworks, leveraging large-scale unlabeled data to enhance prediction accuracy. However, the efficiency of these methods such as Seq3seq, is somewhat limited due to its recurrent neural networks (RNNs) [20] encoder-decoder structure, with the decoder primarily serving as a scaffold during pre-training without significantly contributing to the final prediction or fine-tuning. Hence, these fingerprint methods lack computational efficiency in this regard [12]. In general, these models train deep neural networks to generate robust vector representations using extensive unlabeled datasets. These vector representations are then utilized for supervised training with various classifiers. As these deep models are trained on enormous amount of dataset, the representations are anticipated to contain adequate information for effective inference. However, such methods are not directly trained on prediction tasks, leading to representations optimized solely for the recovery task of the original raw data representation. Consequently, they may not offer optimal inference performance for broader prediction tasks.

Seq2Seq / Seq3Seq fingerprint methods (RNN-based encoder-decoder)

- Architecture: These fingerprint methods such as Seq2Seq typically relies on RNNs, including variations like long short-term memory (LSTM) [20] or gated recurrent units (GRUs) [21], for both the encoder and decoder.
- Encoder: In this approach, the encoder processes an input sequence (e.g., a molecular representation like SMILES) one token at a time. The hidden states are updated sequentially, and the final hidden state summarizes the entire input sequence.
- Decoder: The decoder also uses an RNN and generates the output sequence based on the hidden state from the encoder. In tasks like molecule generation or prediction, the decoder sequentially outputs each element of the sequence (e.g., fingerprint), using the previous output as input for the next.
- Handling long-term dependencies: Since they rely on RNNs, these fingerprint methods such as Seq2Seq can suffer from issues like vanishing gradients when handling long sequences, even with advanced variants like LSTMs. The model must retain the entire sequence context through hidden states passed along each time step.

Transformers [11], by contrast, leverage the self-attention mechanism to model dependencies and interactions within sequences, making them particularly well-suited in this regard for SMILES strings, which encode complex molecular structures in a linear text format. The ability of transformers to process and generate sequences aligns perfectly with the requirements for interpreting SMILES strings, enabling the extraction of rich, contextually aware representations of molecular structures.

Transformer encoder-decoder structure

- Architecture: The Transformer uses self-attention mechanisms instead of RNNs. Both the encoder and decoder are built using layers of multi-head self-attention and feed-forward networks, which operate in parallel rather than sequentially.
- Encoder: The transformer encoder processes the entire input sequence at once, applying self-attention across all tokens in the sequence. This allows the model to capture long-range dependencies and relationships between tokens without the sequential bottleneck of RNNs. Each token attends to every other token in the sequence, capturing context in a global manner.
- Decoder: Similarly, the decoder uses both self-attention and encoder-decoder attention. The self-attention focuses on previously generated tokens (autoregressive generation), while the encoder-decoder attention helps the decoder focus on relevant parts of the input sequence. This parallelization improves efficiency and long-range dependency handling.
- Handling long-term dependencies: Transformers handle long sequences more effectively because self-attention allows every token to directly attend to every other token in the sequence. This avoids the problem of vanishing gradients and limits on sequence length that RNN-based Seq2Seq models may face.
- Attention mechanism (core component): In the transformer, attention mechanisms are the core of both the encoder and decoder. The model doesn't rely on sequential hidden state propagation but instead uses attention to compute dependencies between tokens.

In summary, while both fingerprint methods and transformers follow an encoder-decoder structure, methods such as Seq2Seq or Seq3Seq relies on sequential processing via RNNs, whereas transformers use self-attention mechanisms for parallel processing, allowing them to handle long sequences more effectively and efficiently. Nonetheless, utilizing large amounts of unlabeled SMILES data along with a limited set of labeled data, transformer LMs can effectively learn and capture meaningful molecular representations from SMILES strings [3]. These models first learn from the abundant unlabeled data and are then fine-tuned using labeled data for specific downstream tasks [4]. This approach addresses the challenge of obtaining large quantities of labeled molecular-property data, which is often resource-intensive and requires significant labor through screening experiments and manual data annotation [6,10].

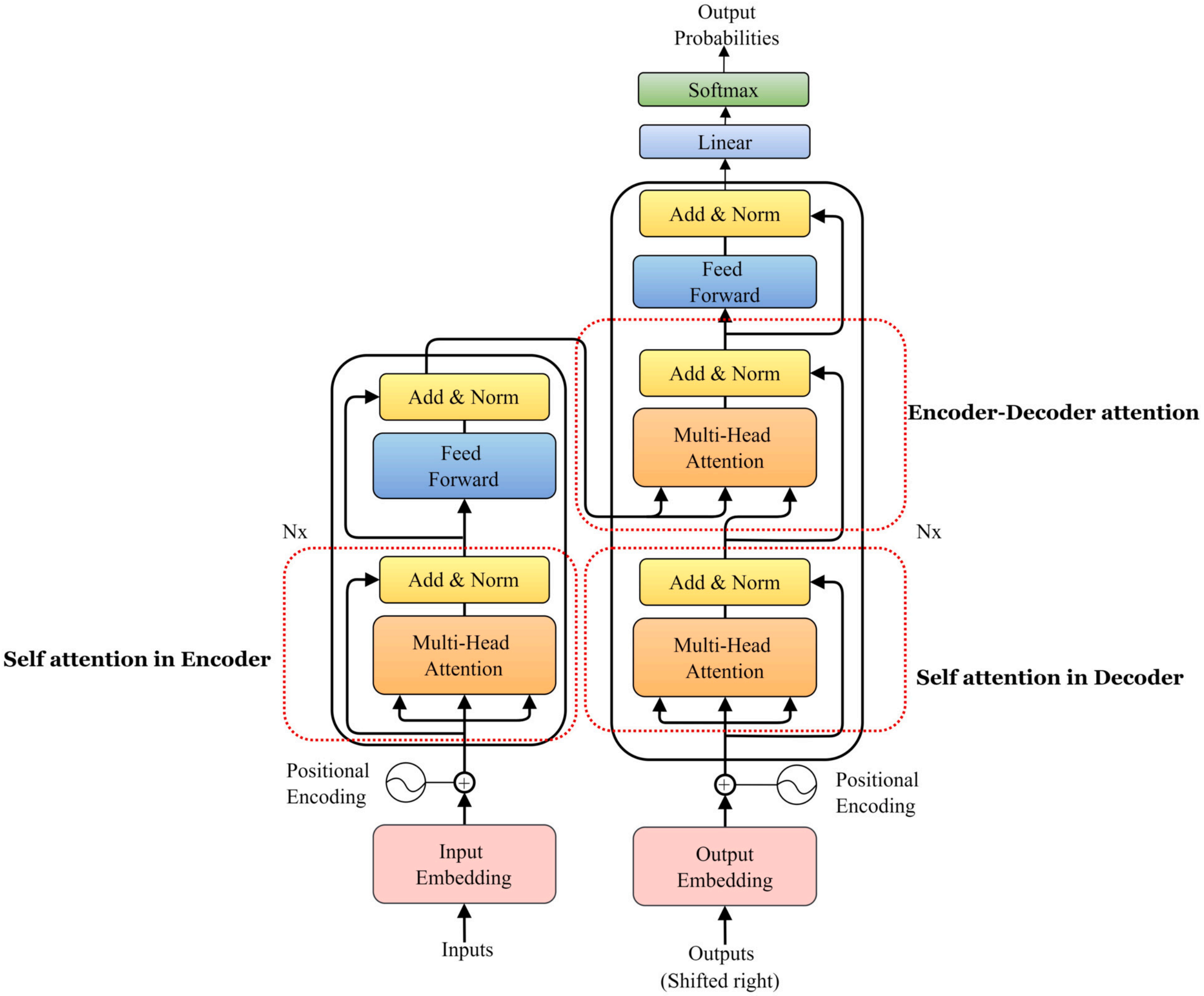

Fig. 1. Transformer (attention-based) architecture. CLMs in this field typically utilize one of three core transformer structures: encoder-only models (e.g., BERT), decoder-only models (e.g., GPT), or the encoder-decoder architecture (e.g., BART, T5).

3. Transformer architecture and attention mechanism

The core structure of a transformer is comprised of two fundamental components: the encoder and the decoder, as illustrated in Fig. 1. Transformer models, originally designed for NLP, have shown remarkable success in various domains, such as computational chemistry. Specific transformer-based architectures, such as bidirectional encoder representations from transformers (BERT) [22] and generative pre-trained transformer (GPT) [23,24] in this regard can be adapted and utilized using for various tasks in cheminformatics such as virtual screening and drug discovery, molecular property prediction, and de novo molecular design. Remarkable capabilities of transformer lies in its innovative self-attention mechanism as shown in Fig. 1. This mechanism, considered the cornerstone of the transformer intelligence, revolutionizes how transformer models capture and understand relationships within sequences [25,26]. Attention mechanisms have proven to be pivotal in capturing long-range dependencies among the elements in an input sequence. Since certain words or SMILES tokens are more impactful in defining the context and meaning of a sequence. The attention mechanism plays a crucial role in assigning the different attention weights to each input elements or tokens [25], The attention mechanism mimics this mechanism by enabling each word (or token) to consider the significance of other words or tokens, thereby grasping their relative importance. The attention weights indicate the significance or relevance of a particular input sequence element at given step, put simply elements with large weight values play a bigger role in comprehending the entire input sequence. This interactive process guarantees that tokens convey contextual information from throughout the entire sequence, which is essential for interpreting complex language structures. Therefore, this facilitates and assists Transformer LMs to concentrate on the most pertinent elements within the given input sequence, aiding in better comprehension and analysis.

The attention mechanism used in Transformer is named scaled dot-product attention for each attention head [11]. It maps or transform the input sequence into three different parts corresponding to query (Q), key (K), and values (V) matrices as depicted in Fig. 2. The terms Q, K, and V utilized in here are inspired by information retrieval systems and databases [25]. It is important to note that, rather than relying on a single attention (single head self-attention) mechanism, multi-head attention utilizes multiple attention mechanisms operating in parallel, enabling the model to capture diverse types of relationships and patterns within the sequence as depicted in Fig. 1. Each separate attention head focuses on various aspects of the input, accelerating and enhancing the attention mechanism's potency [12]. In this context of multi-head attention, three matrices correspond to each single attention head, hence there will be sets of Q, K, and V matrices. Each part (i.e., Q, K, and V) is introduced with weight matrices that can be fit as model parameters during model training, to make self-attention mechanism more flexible and amenable during model optimization for a given input sequence [25]. The Q matrix works together with the K matrix to serve as the input of the Softmax. Then Softmax creates the attention weights, which will be applied to the V matrix to generate the output features with the attention on the whole sequence.

In summary, as shown in Fig. 2, multiple layers of self-attention mechanisms, known as multi-head self-attention repeats the scaled dot-product attention computation multiple times in parallel and the results are concatenated and then linearly transformed, producing a comprehensive and refined representation [26]. Thus the multi heads of attention allow the model to capture information from different parts of the input sequence similar to multiple kernels which produce multiple channels in a convolutional neural network (CNNs) [27], in which each channel capture different feature information. Multi-head attention offers several advantages such as:

- Enhanced feature representation: Multiple attention heads allow the model to focus on different parts of the input sequence simultaneously. Each head can learn to capture different relationships between SMILES tokens, leading to a richer representation of the complex molecular structures.
- Parallelization: Multi-head attention allows multiple attention mechanisms to operate in parallel, which improves computational efficiency compared to processing attention sequentially.
- Improved learning: By using multiple attention heads, the model can better capture various linguistic patterns and dependencies at different levels of abstraction, enhancing its ability to understand both local and global relationships in the complex molecular structures data.
- Reduced overfitting: The use of multiple heads helps mitigate overfitting, as it enables the model to consider diverse perspectives of the same data, rather than relying on a single point of view.

The scaled dot-product self attention is formulated in Equation (1) and shown in Fig. 2.

$$Z = Softmax\left(\frac{(XW^{Q})(XW^{K})^{T}}{\sqrt{d\_k}}\right)XW^{V} \tag{1}$$

where $X \in \mathbb{R}^{N\times M}$ is defined as the input feature matrix, W^Q, W^K, and $W^V \in \mathbb{R}^{M\times dk}$ correspond to the query weight matrix, the key weight matrix, and the value weight matrix respectively. $\sqrt{d_k}$ is a scaling factor preventing softmax from being pushed into a region with an extremely small gradient and Z indicates the output of the attention layer. Fig. 2 illustrates deeper the structural architecture of the attention mechanism.

Attention scores calculation: The self-attention mechanism computes attention scores for each token through the dot product of its Q vector with the K vectors of all other tokens. These scores represent the relevance of the token relative to others in the sequence.

Softmax and Weighted Values: The attention scores are passed through a softmax function, which converts them into probabilities. These probabilities indicate how much weight each token's V vector contributes to the final output.

Output via weighted sum of Vs: The final output for each token is generated by computing the weighted sum of the V vectors, with the weights determined by the softmax probabilities. This output captures both local and global contextual relationships.

Other accompanied components are as explained as follows:

- Normalization layer (Layer Normalization): Stabilizes training by standardizing activations, ensuring they have a mean of zero and variance of one, which helps maintain consistent gradients and prevents exploding or vanishing gradients.
- Feed-forward neural networks: Provide additional non-linearity and learning capacity within each layer, enhancing the model's ability to generalize and capture complex patterns, while the residual connections help in stabilizing this process.
- Residual connections: Skip connections bypass layers, allowing gradients to flow directly through the network, thus mitigating the vanishing gradient problem and helping deeper layers learn more effectively. Each components incorporates a residual input to better utilize the original information or maintain strong information signals to avoiding vanishing gradient phenomenon and overfitting, to ensure stable training.

Together, these elements contribute to efficient training, stability, and generalization in Transformer-based CLMs.

4. Transformer-based models in computational chemistry

Transformer and attention-based models have demonstrated exceptional performance in various tasks in computational chemistry. This section provides an in-depth analysis of how these models contribute to molecular property prediction, virtual screening, and de novo molecular design using SMILES-encoded information.

Molecular Property Prediction: Transformer-based CLMs have shown significant promise in predicting various molecular properties, such as solubility, toxicity, and binding affinity. By training on large datasets of SMILES strings, these models learn to encode

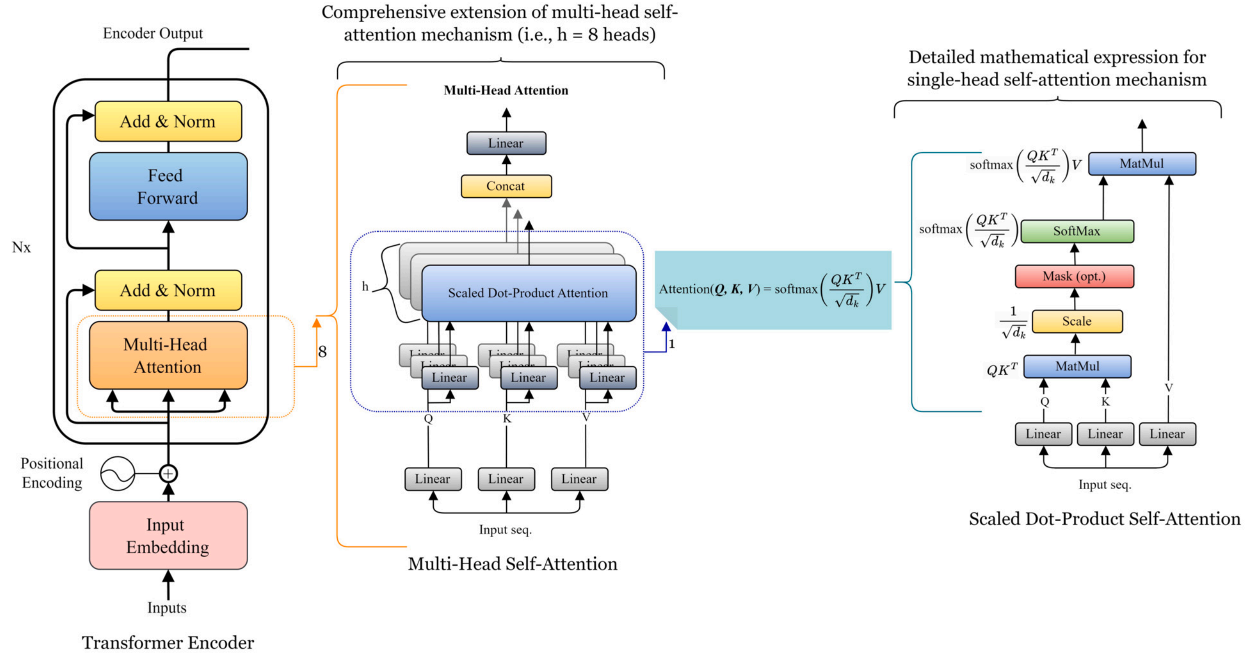
Encoder Output
Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
Inputs
Transformer Encoder
8
Comprehensive extension of multi-head self-attention mechanism (i.e., h = 8 heads)
Multi-Head Attention
Linear
Concat
h
Scaled Dot-Product Attention
Linear
Linear
Linear
Q
K
V
Linear
Linear
Linear
Input seq.
Multi-Head Self-Attention
$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
1
Detailed mathematical expression for single-head self-attention mechanism
$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
MatMul
$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$
SoftMax
Mask (opt.)
$\frac{1}{\sqrt{d_k}}$
Scale
QK^T
MatMul
Q
K
V
Linear
Linear
Linear
Input seq.
Scaled Dot-Product Self-Attention

Fig. 2. Self-attention mechanism architecture. Case study: Encoder.

molecular structures into high-dimensional representations that capture relevant chemical information. Fine-tuning these models on specific prediction tasks further enhances their accuracy, often surpassing traditional machine learning methods.

Virtual Screening and Drug Discovery: In drug discovery, virtual screening is a crucial step where millions of compounds are evaluated for potential biological activity. Transformers can process large chemical libraries encoded in SMILES format, efficiently identifying promising candidates for further experimental validation. The scalability of transformers allows them to handle extensive datasets, making them invaluable in high-throughput screening applications.

De Novo Molecular Design: De novo molecular design involves generating new chemical compounds with desired properties. Transformer models, particularly those with generative capabilities like GPT can be trained to generate valid and novel SMILES strings. This generative process can be guided by desired chemical properties, enabling the design of molecules tailored for specific applications, such as drug development or materials science.

Transformer-based models generally offer several strengths in computational chemistry for molecular properties and chemical reactions analysis using SMILES representation in various criteria such as:

Learn complex relationships: Transformers excel at capturing intricate relationships between elements in sequential data, making them well-suited for analyzing the complex structural patterns present in SMILES strings.

Scalability: These models scale large datasets efficiently, enabling the analysis of extensive molecular libraries containing millions of compounds.

Contextual understanding: Transformers can capture contextual information from the surrounding atoms and bonds, allowing for a more nuanced understanding of molecular structures.

Transfer learning: Pre-trained transformer models can be fine-tuned on molecular property prediction tasks with relatively small amounts of labeled data, leveraging knowledge learned from pre-training on large unlabeled datasets.

4.1. Comparative analysis

Transformer-based models have marked a significant advancement in the field of computational chemistry, particularly in predicting molecular properties from SMILES representations. Table 1 provides a comparative analysis of various transformer-based CLM used in this field. This analysis delves into various categories of transformer models based on the architectures like encoder-based, decoder-based, and encoder-decoder-based, each with unique strength and suited for specific tasks. We further highlighting their applications, strengths, and limitations in molecular property prediction. These models differ in their architectures and specific applications, yet all leverage the core strengths of the transformer architecture as shown in Table 1. For example encoder-based models such as BERT and its variants excel in capturing detailed contextual information or extracting meaningful context from input sequences [26]. These are particularly effective for tasks requiring deep molecular understanding. Decoder-based models like GPT and its derivatives shine in creative generation tasks, enabling the creation of novel molecules with desired properties. Encoder-decoder models combine the strengths of both approaches, offering versatile solutions for complex sequence-to-sequence tasks.

Table 1
Case studies: Transformer-Based Language Models for computational chemistry and cheminformatics.

Backbone	Encoder-based CLMs	Decoder-based CLMs	Encoder-decoder CLMs
Transformers(Attention)	BERT [28] MOLBERT [14] SMILES-BERT [12] ChemBERTa [13] ChemBerta-2 [37] GPT-MolBERTa [40] RoBERTa-based model [43] MoLFormer [45] SELFormer (SELFIES) [47] Mol-BERT [49] MolRoPE-BERT [50] BET [51]	Guo et al. [29] GPT-2-based model [31] MolXPT (GPT-2-based) [33] BioGPT [35] MolGPT (Relative Attention) [38] Mol-Instructions [41]	TransAntivirus (IUPAC) [30] Chemformer (BART-based) [32] MolT5 [34] Transformer-based models [36] X-MOL [39] SMILES transformer [42] Transformer Neural Network [44] Transformer-based ANN [46] RT (SELFIES) [48]

4.1.1. Encoder-based CLMs

- **BERT.** Herein, BERT [28] model was employed as shown in Table 1, and pre-trained on large corpora and fine-tuned for specific tasks. In cheminformatics, it has been adapted to process SMILES strings, enabling the extraction of rich, contextual molecular representations. Applications: Property prediction, molecular similarity search, and drug discovery. Strengths: BERT captures bidirectional context, enhancing the understanding of chemical environments and molecular interactions. They excel in transfer learning, requiring smaller datasets for fine-tuning. Limitations: High computational cost and memory usage due to the bidirectional nature of the model.
- **MOLBERT.** This model is a chemistry-specific adaptation of BERT, tailored for handling SMILES strings. Applications: Predicting physicochemical properties, bioactivity, and molecular interactions. Strengths: Specialized training on chemical datasets enhances performance on cheminformatics tasks. Limitations: Requires large, labeled datasets for optimal performance.
- **SMILES-BERT.** This is a variant of BERT explicitly designed to work with SMILES representations. Applications: Predicting molecular properties and chemical reactions. Strengths: Effective at learning molecular representations directly from SMILES strings without requiring extensive feature engineering. Limitations: Like other BERT-based models, it demands substantial computational resources.
- **ChemBERTa and ChemBERTa-2.** These models are based on the RoBERTa [52] which is the adaptation of BERT [22], designed for chemical property prediction, optimized for handling SMILES inputs. Applications: Broad range of property predictions, including toxicity, solubility, and bioactivity. Strengths: Enhanced with domain-specific training, improving prediction accuracy. Limitations: The complexity and resource demands are significant.
- **RoBERTa-based Model.** This model also refine BERT by training with more data and longer sequences, further adapted for cheminformatics. Applications: Property prediction and molecular classification tasks. Strengths: Improved training protocols lead to better generalization in chemical datasets. Limitations: Increased computational requirements for training and inference.
- **Mol-BERT and MolRoPE-BERT.** These models are also based on and utilized BERT architecture to handle chemical SMILES representation for predicting molecular properties and chemical reactions. The primary distinction between these models lies in their approach to position embedding (PE) implementation. Mol-BERT [49] model employs absolute PE, which limits the length of encoding sequences, impacting the effectiveness of downstream model prediction tasks [50]. On the other hand to address the aforementioned issues, Liu et al. [50] proposed MolRoPE-BERT, which applies rotary PE.

4.1.2. Decoder-based CLMs

- **GPT-2-based Models (e.g., MolGPT, MolXPT).** In overview these models, derived from the generative pre-trained transformer 2 (GPT-2) [24], are used for generative tasks, including SMILES string generation. Applications: De novo drug design and molecular optimization. Strengths: Excellent at generating novel, chemically valid molecules by leveraging learned distributions of SMILES strings. Limitations: Tend to focus on one-directional context, which may limit the understanding of bidirectional dependencies in molecules.
- **BioGPT.** This model focuses on generating biomedical text and mining. It appears to outperform other related models such as BioBERT [53], PubMedBERT [54], GPT-2 model in this regard. Applications: Biomedical text generation and Mining. Strength: Tailored to handle large-scale biological data, making it suitable for biomedicine, biomedical text mining and generation. Limitation: Generative focus may not always capture the most relevant features for certain predictive tasks.
- **MolGPT.** This proposed model employs the generative pre-trained transformer (GPT) architecture along with relative attention for de novo drug design [38]. Relative attention, a variation of the attention mechanism, enables the model to understand the relative distances and relationships between tokens in the input sequence. Unlike the standard attention mechanism, which uses fixed-PEs to encode positional information, relative attention dynamically incorporates relative positional encodings during the attention calculation. This enables the model to learn the syntax of new and unseen tokens efficiently.

4.1.3. Encoder-decoder-based CLMs

- **TransAntivirus (IUPAC).** A model designed to predict antiviral properties by processing chemical structures in the IUPAC nomenclature format. Applications: Antiviral drug discovery. Strengths: Specifically optimized for antiviral property prediction, leveraging encoder-decoder architecture for better sequence-to-sequence learning. Limitations: Performance may be dataset-dependent, requiring high-quality training data.

- **Chemformer (BART-based).** Based on the Bidirectional and Auto-Regressive Transformers (BART) architecture, Chemformer is tailored for generating and predicting molecular properties. Applications: Molecular property prediction and compound generation. Strengths: Combines the strengths of both encoder and decoder models, providing robust performance in property prediction and generation tasks. Limitations: Complex training process and higher computational overhead.
- **SMILES Transformer.** This model utilizes the transformer architecture for converting SMILES strings into meaningful representations and predicting their properties. Applications: Broad applications including drug discovery, toxicity prediction, and property optimization. Strengths: Effective at handling the sequential nature of SMILES strings, capturing intricate molecular features. Limitations: Computationally intensive and requires significant preprocessing.
- **MolT5.** An adaptation of the T5 (Text-to-Text Transfer Transformer) model for molecular tasks. Applications: Molecular property prediction, reaction prediction, and molecular optimization. Strengths: Unified framework for various tasks, leveraging transfer learning for improved performance. Limitations: Training and fine-tuning require substantial computational resources and high-quality data.

In general, transformer-based CLMs especially encoder-based models have been widely employed and contributed significantly in this field followed by encoder-decoder based models with few studies utilize molecular representations other than SMILES such as SELFIES, IUPAC, and biomedicine text data representations as shown in Table 1.

5. Materials and methods

5.1. Datasets

Dataset in this field in primarily various chemical compound representations as shown in Table 4. In this review, most of LMs primarily utilized SMILES dataset however few studies have used IUPAC and SELFIES as shown in Table 1. DeepSMILES is defined as a variant of SMILES, altering how branches and rings are represented while maintaining the same fundamental atom-level characters with SMILES [55]. SMILES is a notation that allows a user to represent a chemical structure in a way that can be used by the computer. A SMILES training dataset would typically consist of molecular structures encoded using the SMILES notation, paired with corresponding labels or information about the properties of the molecules. SMILES strings are composed of ASCII characters and provide a textual representation of molecular structures. Ensuring the accuracy or correctness of SMILES strings representations is crucial, as invalid representations can result in the misinterpretation of molecular structures by the model. Data preprocessing may involve tasks such as canonicalization of SMILES strings (ensuring a unique representation for the same molecule), handling chirality, and other relevant transformations to standardize the data. Creating a SMILES training dataset often involves leveraging existing databases of chemical compounds, experimental data, or cheminformatics resources such as ZINC [5,56], PubChem [57,58], and ChEMBL [59,60] databases. Researchers and practitioners in computational chemistry or cheminformatics use such datasets to develop models for tasks such as molecular property prediction, drug discovery, and chemical informatics.

5.1.1. Pre-training dataset

The dataset serves as the input to the ML algorithm, facilitating the learning of patterns, relationships, and features by the LMs. Herein, dataset is categorized into two primary groups based on pre-training and fine-tuning strategies, aiming to fully utilize the potential of transformer-based CLMs. A pre-training dataset typically enables the model to acquire more robust patterns and enhance generalization to unseen data. Creating an effective pre-training dataset involves meticulous data collection, pre-processing, and cleaning to ensure that the model learns meaningful patterns. The quality of the pre-training dataset significantly impacts the model's performance and it's ability to generalize to new data.

5.1.2. Fine-tuning dataset

The fine-tuning serves as the foundation for adapting a pre-trained model to a specific task or domain, enabling the model to learn task-specific patterns and achieve optimal performance, leveraging transfer learning by building on the knowledge and representations learned by the pre-trained model. The fine-tuning dataset plays a crucial role in this process by providing examples that are relevant to the target task. The size of the fine-tuning dataset can vary depending on factors such as task complexity, the availability of labeled data, and computational resources. Generally, a larger and higher quality fine-tuning dataset with accurate labels and minimal errors, enhances the model's capacity to learn robust patterns and generalize effectively to unseen data. Conversely, poor-quality data may impede performance and necessitate additional pre-processing or cleaning efforts. Fine-tuning datasets are often divided into training, validation, and test sets, with the former employed to update the model's parameters during fine-tuning and the latter utilized to monitor performance and fine-tune hyperparameters.

All transformer-based CLMs in this review study are fine-tuned or evaluated on several regression and classification tasks from MoleculeNet introduced by Wu et al. [61]. MoleculeNet, a large scale benchmark for molecular machine learning (ML) that curates multiple public datasets as shown in Table 2. However most data still struggle to deal with complex tasks under data scarcity and highly imbalanced classification.

Table 2
Datasets used in cheminformatics for pre-training and fine-tuning transformer-based CLMs.

Pre-training		Fine-tuning		SMILES Data	Task	# Task	# Compounds
Databases	ZINC [5,56] PubChem [57,58] ChEMBL [59,60]	Category	Physical Chemistry [61]	ESOL	R	1	1128
				FreeSolv	R	1	642
				Lipophilicity	R	1	4200
			Biophysics [61]	PCBA	C	128	437929
				MUV	C	17	93087
				HIV	C	1	41127
				PDBbind	R	1	11908
				BACE	C	1	15513
			Physiology [61]	BBBP	C	1	2039
				Tox21	C	12	7831
				ToxCast	C	617	8575
				SIDER	C	27	1427
				ClinTox	C	2	1478
			Proposed [62–64]	Antimalarial	C	1	4794
				Cocrystals	C	1	3282
				Covid	C	1	740
				Genes	R	1	201

R - Regression.
C - Classification.

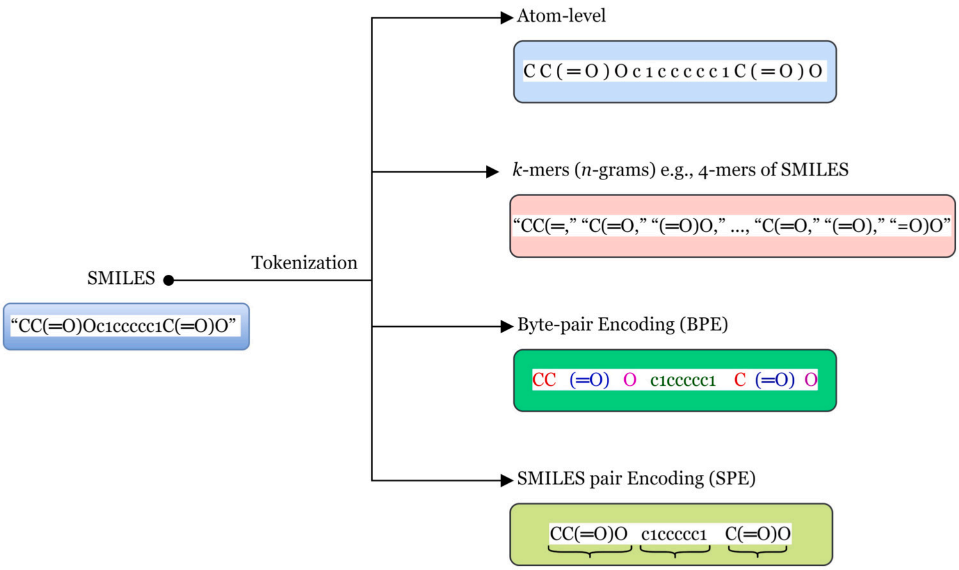
Atom-level
C C (= O) O c 1 c c c c c 1 C (= O) O
k-mers (n-grams) e.g., 4-mers of SMILES
"CC(=," "C(=O," "(=O)O," ..., "C(=O," "(=O)," "=O)O"
SMILES
"CC(=O)Oc1ccccc1C(=O)O"
Tokenization
Byte-pair Encoding (BPE)
CC (=O) O c1ccccc1 C (=O) O
SMILES pair Encoding (SPE)
CC(=O)O c1ccccc1 C(=O)O

Fig. 3. Tokenization mechanism.

5.2. SMILES tokenization and encoding

In the context of NLP especially transformer-based models in cheminformatics, tokenization refers to the process of breaking down SMILES strings data into their constituent elements or substructure tokens. Tokens are then converted into embeddings—numerical representations that capture the semantic information of the tokens. These embeddings undergo processing through a series of identical layers, with each layer comprising two sub-layers: multi-head self-attention and position-wise feedforward neural networks. SMILES strings represents chemical structures using a simplified notation system, where each character or group of characters represents specific elements, bonds, or substructures features such as functional groups. During data-driven tokenization of SMILES data, each character or group of characters that represents a specific element, bond, or structural feature is identified and treated as a separate token as shown in Fig. 3. The partitioning or tokenization alter the perception of molecules owing to the internal relationships among its atomic elements or components, and allows NLP models to process and understand the chemical structure represented by the SMILES strings data more effectively. Tokenization is a crucial preprocessing step before feeding SMILES data into a NLP models to facilitate both pretraining and fine-tuning tasks such as chemical property prediction, reaction prediction (i.e., drug-target predictions, API-coformer predictions). Tokenization can significantly influence prediction quality within the framework of text generation [65]. In this realm, tokenization of SMILES accelerates and increases the SMILES-based prediction performance [66].

In the feld of cheminformatics, SMILES tokenization is typically different compared to other NLP domains which involve features like white spaces within a sequence and more than one sequence of sentences. In SMILES, there is no presence of white space, and

Table 3
Tokenization strategies from Previous studies.

Study	Tokenization algorithm
Ucak et al. [65]	Atom-in-SMILES (AIS)
Li et al. [67]	SMILES pair encoding (SPE) imitating BPE atom-level tokenization k-mer (also known as n-grams) tokenization
Chithrananda et al. [13]	SMILESTokenizer Byte-Pair Encoder (BPE)
Lee et al. [28]	Byte-Pair Encoder (BPE)

parentheses play the essential role of grammatical punctuation by indicating the start and end of a branch [28]. Therefore, square brackets are used to indicate atoms and charged atoms. During tokenization parentheses and square brackets are crucial for tokenizers to understand the substructures effectively. Additionally, there are two tokens namely < CLS > defined as the class token and < SEP > which indicates separation token, however in SMILES case it represent end of sequence as there is only single sequence input. After tokenization of SMILES strings, each token is mapped to three vectors: Q, K, and V, which are generated from the embeddings of the input tokens.

Previous studies have discussed and demonstrated various tokenization algorithms of SMILES towards SMILES-based prediction performance as shown in Table 3 and Fig. 3. In the field of Chemistry, tokenizing SMILES into atom-wise tokens is primarily method used for training chemical NLP models utilizes the character-level tokenization with certain adjustment to ensure that atoms are identified and extracted as individual tokens. Alongside this atom-wise SMILES tokenization approach, there have been recent introductions of novel molecular representations like SELFIES and DeepSMILES. Additionally, specialized tokenization methods such as byte-pair encoding (BPE), have emerged [65]. BPE is a hybrid combines elements of both character and word-level representations, enabling the management of extensive vocabularies in natural language collections. It operates on the principle that uncommon or unfamiliar words can often be broken down into recognizable subwords. BPE identifies the optimal segmentation of words by iteratively and greedily combining commonly occurring pairs of characters [13]. As observed in Table 3, most of the studies have been inspired or employed BPE as their architectural back-bone for their proposed tokenization algorithms such as a data-driven substructural tokenization algorithm namely SMILES pair encoding (SPE) [67].

5.3. Embedding information

Following the tokenization of the SMILES string data is the computation of embedding information. The tokens are first embedded into the feature space. Besides the token embedding, positional embedding is also included to add sequential information used in self-attention layer to utilize the temporal information of the inputs. Positional embeddings (PEs) are crucial for representing sequential data and grasping complex relationships within sequences, allowing models to capture both the presence and the order of tokens. Various PEs have been utilized such as absolute, relative key, relative key query, rotary [68], and sinusoidal positional encoding [11]

Notably, transformer-based CLMs in chemistry differs from other domains in the following perspectives:

- Single Masked SMILES Recovery is employed on large scale unlabeled dataset.
- Segmentation embedding is not included in models like BERT and their variants since SMILES data does not involve the continuous sentences training.

5.4. Pre-training

In the context of transformer-based language models (LMs) applied to SMILES, pre-training involves training the model on a large dataset of chemical structures represented using SMILES notation. The model architecture used for pre-training is based on transformers, a type of neural network (NN) architecture known for its self-attention mechanism which is particularly well-suited for learning dependencies. As a result, transformer-based LMs have demonstrated great effectiveness in capturing long-range dependencies and contextual information in sequences or within SMILES strings. Pre-training in transformer-based CLMs refers to the initial phase of training where a model is pre-trained on a large corpus and diverse corpus of unlabeled text data (i.e., SMILES) before being fine-tuned on a specific downstream tasks. Thus, pretraining offers several advantages, especially in chemical tasks such as:

- Rich Representations: Pre-training on large datasets (e.g., molecular structures or chemical reactions) helps the model learn general patterns, such as chemical bonding, molecular interactions, and reaction mechanisms. This provides a strong foundation for downstream tasks.
- Data Efficiency: Pre-training reduces the need for large, labeled datasets in specific chemical tasks, as the model can leverage knowledge from pre-training to quickly adapt to new tasks with fewer data points.
- Improved Generalization: Pre-trained models generalize better, as they have been exposed to diverse chemical data. This improves their ability to handle complex tasks like molecular property prediction or reaction outcome forecasting.

Herein, the CLMs are trained on a large amount of data without any specific task labels, and this is often referred to as an unsupervised learning strategy. During unsupervised pre-training, the objective is typically a language modeling task known as masked language modeling (MLM), where the model learns to predict the next word or token in a sequence given the context of the preceding words. Pre-training allows the model to capture rich contextual information, general language patterns, semantics and linguistic nuances from diverse text sources in this case chemical contextual and topological information of SMILES compounds. The pre-trained model acts as a strong language understanding base that can be adapted to different tasks, making it more effective and efficient in which it has been successful in improving the performance of language models (LMs). The size of the pre-trained model and the amount of data are crucial factors in determining the model's performance.

5.5. Fine-tuning

Fine-tuning is the subsequent phase after pre-training, where the learned parameters of the pre-trained model can be transferred to a specific downstream task. This fine-tuning involves further training the model on a smaller, task-specific labeled datasets, adjusting the weights based on the task's objectives and more focused on the specific application or domain of interest. The pre-trained model, having learned general language features and equipped with a foundational understanding of the structural characteristics and chemical context of molecules encoded in SMILES notation, can then be fine-tuned on task-specific datasets for more specialized applications within the domain of computational chemistry and cheminformatics. Fine-tuning is faster and requires less labeled data than training a model from scratch for various tasks such as classification and regression. Therefore, fine-tuning is paramount considering various factors such as:

- Task Specialization: Chemical tasks (e.g., toxicity prediction, drug discovery) often require specific knowledge. Fine-tuning adapts the general representations from pre-training to these specialized tasks, allowing the model to focus on relevant features.
- Performance Boost: Fine-tuning refines the model's parameters on the specific task data, leading to improved accuracy and task-specific performance, ensuring that the pre-trained knowledge is fully leveraged and optimized for the target chemical problem.

Thus, pre-training provides a broad knowledge base for the model to learn a general and broad understanding of the relationships, patterns, and features within molecular structures encoded in the SMILES format. While fine-tuning adapts this general knowledge to specific cheminformatics downstream tasks, leading to more effective and task-specific models. This two-step process has been instrumental in the success of transformer-based CLMs in various NLP applications. In addition, to enable meaningful comparisons between CLMs in cheminformatics, a standardized set of evaluation metrics is essential. These metrics assess the performance across multiple dimensions, such as molecular representation, property prediction, and reaction modeling. These can be observed in various works in cheminformatics including Chemberta [13], SMILES-BERT [12], Mol-BERT [49].

6. Challenges and limitations

This section addresses these challenges and suggests potential avenues for future research. Although successful, transformer and attention-based models face limitations in handling large-scale chemical datasets.

Data efficiency: Despite their ability to generalize well with limited labeled data, transformer models may still require substantial amounts of training data to achieve optimal performance, particularly for complex tasks or rare molecular properties. Given that pre-training language models such as CLMs requires substantial amounts of data, this challenge can be addressed through various strategies to mitigate the demand for extensive datasets:

- Transfer Learning: By pre-training models on large, general chemical datasets, then fine-tuning on specific tasks with limited data, the model leverages learned knowledge to improve performance on data-scarce tasks.
- Data Augmentation: This involves creating variations of existing chemical data, such as altering molecular structures while preserving their properties, to increase the diversity of training samples without needing new data.
- Synthetic Data Generation: Generating synthetic chemical data using techniques like generative models such as generative adversarial networks (GANs) [69] or variational autoencoders (VAEs) [70] can provide additional training examples, enhancing data efficiency and helping the model learn a broader set of chemical patterns.

These approaches improve data efficiency, reduce dependency on large datasets, and enhance the performance of chemical language models.

Interpretability: The inherent complexity of transformer architectures can make it challenging to interpret the learned representations and understand how specific molecular features contribute to property predictions.

Computational resources: Training transformer models, especially large-scale variants, demands significant computational resources and time, which may limit their accessibility to researchers with constrained resources. Computational resource constraints are a significant concern when training and deploying large chemical language models (CLMs). To address these limitations, various optimization techniques can be employed:

Table 4
Molecular strings representations case study: ibuprofen.

Type	Molecules representation
IUPAC Name	2-[4-(2-methylpropyl)phenyl]propanoic acid
SMILES (Canonical or Isomeric)	CC(C)Cc1ccc(cc1)[C@@H](C)C(=O)O
DeepSMILES [55]	CCC)Cccccc6))[C@@H]C)C==O)O
SELFIES [74]	[C][C][Branch1][C][C][C][C][=C][C][=C][Branch1][Branch1][C][=C][Ring1][=Branch1][C@@H1][Branch1][C][C][C][=Branch1][C][=O][O]
InChI	1S/C13H18O2/c1-9(2)8-11-4-6-12(7-5-11)10(3)13(14)15/h4-7,9–10H,8H2,1–3H3,(H,14,15)
InChIKey	HEFNNWSXXWATRW-UHFFFAOYSA-N

- Model Pruning: This involves removing less critical neurons or layers in the model, reducing its size and computational demands while maintaining performance.
- Quantization: By reducing the precision of model weights (e.g., using 8-bit instead of 32-bit floating points), quantization lowers memory and computation requirements, enabling faster inference without significant loss in accuracy.
- Distributed Training: Distributing the training process across multiple GPUs or nodes accelerates the training of large models, allowing them to scale efficiently and complete in a reasonable time frame despite hardware constraints.

These strategies help optimize CLMs for efficiency, making them more accessible in resource-limited environments.

Handling rare molecules: Transformers may struggle to effectively model rare or novel molecular structures that deviate significantly from the patterns observed in the training data. Handling rare molecules in CLMs is challenging due to the limited availability of training examples for these compounds. This might be addressed through:

- Domain Adaptation: Adapting a pre-trained model from a broad chemical domain to a specific subset of rare molecules helps the model better capture unique molecular features, improving performance even with limited examples.
- Few-Shot Learning: This technique allows CLMs to learn from only a few examples by leveraging prior knowledge from pre-training. It enables the model to generalize to unseen or rare molecules with minimal additional data.

Both strategies help improve the model's ability to handle rare molecules by efficiently transferring or adapting knowledge from more common data.

Majority of the attention-based CLMs proposed and discussed in this review utilize SMILES particularly canonical SMILES notations, which serves as the most widely adopted string-based encoding for molecules, in both pre-training and fine-tuning phase in this field as shown in Table 1. Despite being widely used and showing potential utility in computational chemistry, SMILES has its limitations and poses several challenges concerning its validity and robustness, potentially impeding the model's ability to effectively extract the underlying knowledge within the data [47]. As discussed by Yüksel et al. [47] the limitations of SMILES-based representations include:

- The potential for non-canonical representations of the same molecule, reducing string uniqueness.
- In SMILES notation, many symbol combinations yield invalid outcomes that lack a valid chemical interpretation [71].
- The possibility of valid SMILES strings having invalid chemical properties, such as exceeding atom valency [72].
- Inadequate capture of spatial information.
- Lastly insufficiency in fully representing molecular characteristics due to its lack of syntactic and semantic robustness [71,73].

Furthermore, current benchmark datasets may lack the comprehensiveness needed to tackle emerging real-world challenges and diseases effectively. Therefore, the integration of novel benchmark datasets focusing on diseases like malaria, cancer, and COVID-19 is essential. These new datasets can significantly aid transformer-based CLMs in comprehensively exploring and evaluating their efficacy in addressing complex real-world scenarios. By incorporating datasets that specifically target these diseases, researchers can enhance the capacity of transformer-based models to address pressing healthcare challenges and contribute to the advancement of diagnostic and therapeutic solutions.

7. Future research

Novel molecular representation: While SMILES stands as the most commonly and widely used and has demonstrated potential utility in computational chemistry, it is somehow insufficient to fully represent the complexity and properties of the molecules [47] Therefore, other molecular representations as shown in Table 4 such as DeepSMILES, SELFIES, IUPAC Names, InChI could be taken into consideration and utilized during pre-training and fine-tuning to access the performance of transformer and attention-based models. For example, DeepSMILES were proposed as an improvement of SMILES, to address unbalanced parentheses and ring closure pairs which cause invalid syntax [55]. SELFIES encapsulates all molecular graph information, including atoms, bonds, branches, and rings using characters enclosed in brackets and therefore, it appears to effectively resemble the unique identification of a given molecule [71,74]. For example, normal benzene SMILES is ‘c1ccccc1’ when converted from SMILES to SELFIES will be ‘[C][=C][C][=C][C][=C][Ring1][=Branch1]’, and lastly if converted from SELFIES back to SMILES will be ‘C1=CC=CC=C1’. Recent discussions have extensively explored potential future extensions of SELFIES aimed at addressing the limitations of current molecular string representations [3,71]. In addition SELFIES may also serve as inspiration for potential variants of SMILES and DeepSMILES [3]. Expanding chemical languages to encompass more intricate molecular entities presents a promising avenue for advancing the capabilities of transformer based CLMs in chemistry. This expansion could involve addressing complex structures like proteins and peptides containing non-natural amino acids, as well as exploring areas such as crystals and supramolecular chemistry.

Role of SMILES token characters: Understanding the role of different SMILES token characters—such as metals, non-metals, bonds, and branches—is crucial in the development and enhancement of chemical language models (CLMs). By examining how these specific parts of the SMILES strings impact the model’s performance, especially in terms of information loss, researchers can improve the focus of CLMs on essential tokens rather than relying on random masked language modeling (MLM). For instance, during the pre-training phase of MLM, researchers can strategically hide tokens representing metals, non-metals, bonds, or branches in the SMILES sequences. By doing so, they can assess how the model performs with these tokens masked, thereby gaining insights into which components are most critical for accurate molecular property prediction. This targeted masking approach helps in refining the CLMs to concentrate on significant elements of the SMILES strings, ultimately enhancing their predictive capabilities.

Few-shot learning: The approach of few-shot learning is particularly pertinent in fields like NLP and Chem-informatics, where acquiring labeled data can be challenging. Few-shot learning entails training a model with a small number of labeled examples per class, aiming for effective generalization with minimal training data. This approach becomes particularly valuable in situations where gathering extensive labeled data is impractical or costly. By combining few-shot learning approaches with large-scale pre-trained transformer based CLMs, prospective applications of CLMs are anticipated to see significant enhancements [3,75]. Furthermore, enhancing the capability of CLMs in cheminformatics to propose synthesizable molecules holds the potential to elevate their practical relevance in both cheminformatics and bionformatics for drug discovery efforts [76].

Effective analysis in drug repurposing or repositioning: Drug repurposing, also known as drug repositioning, represents a cost-effective strategy for identifying new therapeutic indications for existing drugs within a short period of time. This approach holds promise in addressing challenges such as the emergence of drug resistance and the reemerging of viral infections [77]. In the context of drug repurposing, certain drug molecules may exhibit diverse biological activities, showing efficacy or lack thereof against specific targeted organisms across different clinical trials aimed at treating various diseases using similar drug compounds. For instance, a chemical compound that demonstrates promising efficacy in treating COVID-19 may exhibit inactive effects when used to combat malaria. Therefore, it becomes imperative for LMs such as transformer based models to discern between identical compound molecules exhibiting multiple biological or chemical activity profiles across different therapeutic domains.

Improvement of donwnstream task datasets: Recent benchmark datasets [61] in the field of NLP shown in Table 2, while valuable for assessing model performance, may not fully address the complexities and challenges encountered in real-world scenarios, particularly in the context of diseases like malaria, cancer, and COVID-19. To overcome this limitation and facilitate more comprehensive evaluations, it is imperative to incorporate new benchmark datasets that specifically target these diseases. By leveraging datasets that reflect the intricacies of these conditions, transformer-based models can be better equipped to explore and evaluate their effectiveness in addressing real-world challenges. This approach not only enhances the applicability of these models but also contributes to advancements in disease diagnosis, treatment, and prevention strategies. Therefore, in this review, below and as shown in Table 2, we propose various new datasets that we are convinced can be added and utilized during fine-tuning to address the aforementioned challenges.

1. Covid dataset [78]. This dataset consist of drug chemical compounds that have shown experimental activeness against coronavirus drug targets such as SARS-CoV, MERS-CoV, SARS-CoV-2, ORF1ab - ORF1a polyprotein; ORF1ab polyprotein (Betacoronavirus England 1), surface glycoprotein (SARSCoV-2), and Replicase polyprotein 1ab (SARS-CoV-2)). Furthermore, in this dataset we incorporate the information about the chemical structure representation of the proven drug target enzyme i.e., Mpro, or 3CLpro which is a proven drug discovery target in the case of the SARS-CoV-2 that shares a genome sequence similar to that of other members of the betacoronavirus group like MERS-CoV and SARS-CoV. This target is also essential for the replication and life cycle of coronaviruses.
2. Cocrystal formation dataset [62]. The dataset comprises chemical reactions between active pharmaceutical ingredients (APIs) and different coformers, with the aim of forming cocrystals. APIs have garnered significant pharmaceutical interest due to their intrinsic properties, and are consequently employed as adjunct solid dosage forms following FDA guidance and approval of pharmaceutical cocrystals as a promising avenue for drug substance development.
3. Antimalarial drugs dataset [63]. This dataset encompasses drugs exhibiting antimalarial activity against Plasmodium falciparum, serving as a valuable resource for various ML models aimed at predicting and facilitating the discovery of novel antimalarial drugs. In addition to drugs, the dataset includes three potential and high-priority targets—Acetyl CoA Synthetase (PfAcAS), Bifunctional Farnesyl/Geranylgeranyl Pyrophosphate Synthase (F/GGPPS), and Monoacylglycerol Lipase (PfMAGL)—present in Plasmodium and crucial for advancing antimalarial drug discovery efforts. Despite global initiatives, malaria remains a formidable health challenge, affecting nearly half of the world’s population in 2020. The emergence of parasite resistance to existing antimalarial

treatments or drugs have drawn the attention and critical need for innovative therapeutic strategies capable of addressing the diverse stages of the Plasmodium life cycle.

4. Cancer dataset [64]. The dataset comprises a gene expression matrix containing data from 8,046 genes across 734 cell lines, along with a drug response matrix encompassing 201 drugs tested on the same cell lines. By integrating gene expression data with biological networks, the dataset encapsulates crucial insights into the mechanisms underlying drug responses in cancer treatment. Cancer patients' genomic profiles, including gene expression patterns, have emerged as pivotal resources for predicting drug responses in the realm of personalized medicine. With the availability of extensive drug screening datasets involving cancer cell lines, numerous computational methodologies, notably ML approaches, have been employed for the prediction of drug responses.

However, when fine-tuning transformer-based CLMs, it is crucial that newly proposed datasets meet certain criteria to ensure the model learns to generalize across diverse chemical tasks and environments. Such key considerations include:

- Diverse chemical structures: The dataset should encompass a wide range of molecular structures, including small molecules, polymers, peptides, and more complex organic and inorganic compounds. This ensures that the model is exposed to different bonding patterns, functional groups, and stereochemistry, enhancing its ability to understand a variety of molecular architectures.
- Varying levels of complexity: Including molecules with different levels of structural and chemical complexity is essential. This means incorporating both simple, well-known compounds (e.g., methane, ethanol) and more intricate ones (e.g., natural products, pharmaceuticals). This allows the model to develop a robust understanding of both fundamental chemical principles and more nuanced molecular behaviors.
- Representation of different chemical environments: The dataset should capture molecules under various conditions, such as different solvents, temperatures, pH levels, and reaction environments. This helps the model adapt to the variability in how molecules behave in different settings, which is critical for tasks like reaction prediction and property estimation.
- Coverage of functional groups and reactivity: A well-rounded dataset should include molecules with a wide range of functional groups (e.g., alcohols, amines, ketones) and reactivity patterns. This diversity ensures that the model can predict and understand chemical transformations and interactions relevant to different fields, from organic synthesis to material science.
- Balanced data representation: Ensuring that no single class of molecules dominates the dataset is crucial for balanced learning. Over-representation of certain types of molecules (e.g., small organic compounds) could bias the model and hinder its performance on other less-represented molecular classes (e.g., transition metal complexes).
- Task-relevant annotations: For specific fine-tuning tasks, the dataset should contain relevant annotations, such as molecular properties (e.g., solubility, toxicity, reactivity), bioactivity data, or reaction outcomes. This allows the CLM to specialize in predictive tasks or property estimation by learning the associations between chemical structures and their real-world behaviors.

By ensuring that fine-tuning datasets meet these criteria, the chemical language model will be able to generalize well to a broad range of tasks, from molecular property prediction to chemical reaction modeling, and improve its ability to handle the complexities of real-world chemical environments.

The dual use risk of and potential harms of CLMs: The integration of AI such as CLMs in chemistry indeed raises significant ethical questions, particularly regarding unintended consequences and biases. For instance, Ahmad et al. [37] not only open-sourced their trained models but also emphasized the dual-use risks associated with CLMs. They argue that the challenge of synthesizing novel molecules serves as a safeguard against potential harms that could arise from the release of updated models. Furthermore, they acknowledge that the balance of risks may change as their research progresses, and they have committed to continuously evaluating the dual-use implications of future open-source releases.

- **Unintended consequences**
  - Safety risks: CLMs may suggest novel chemical compounds or reactions that, while theoretically sound, could pose safety risks in practice. Unanticipated reactions or toxic byproducts might emerge, leading to hazardous outcomes.
  - Environmental impact: CLM-driven processes for materials synthesis or drug discovery might inadvertently prioritize efficiency over sustainability, resulting in harmful environmental effects. This includes the potential for increased waste or the development of substances that could harm ecosystems.
  - Over-reliance on automation: As CLMs, take on more decision-making roles, there may be a risk of diminishing critical thinking among researchers, leading to over-reliance on AI suggestions without adequate scrutiny or validation.

- **Biases**
  - Data Bias: CLMs are trained on existing datasets, which may be biased due to underrepresentation of certain chemical structures or properties. This can lead to skewed predictions that favor more common compounds while neglecting rare or novel ones, limiting innovation and discovery in important areas.
  - Reinforcement of existing inequalities: If the datasets used to train CLMs reflect historical biases in chemical research (such as underrepresentation of certain demographics in drug development), the CLMs may perpetuate these inequalities by favoring the needs of well-studied populations while overlooking others.

- Interpretation bias: The interpretation of CLM-generated results may be influenced by human biases, affecting how findings are perceived and applied in research and industry. This could lead to misinformed decisions based on CMLs outputs.

Addressing these ethical questions requires a proactive approach, including careful dataset curation, ongoing evaluation of CLMs, and fostering a culture of responsible CLMs use in the chemical sciences. Engaging with diverse stakeholders and prioritizing transparency can help mitigate risks and ensure that CLMs serves to advance ethical and equitable practices in chemistry.

8. Discussion

Dataset bias can significantly impact the performance of CLMs, potentially limiting their generalizability and effectiveness in real-world applications. When datasets are not representative of the broad spectrum of chemical diversity and complexity, models trained on them may develop skewed or incomplete understandings of chemical principles. Here are the potential impacts of dataset bias and why ensuring representative datasets is crucial:

- Reduced generalization: If a dataset is biased toward specific classes of molecules (e.g., small organic compounds), the CLM may struggle to generalize to underrepresented molecular categories, such as inorganic complexes, natural products, or polymers. This leads to poor performance when the model encounters unfamiliar chemical structures or environments in real-world tasks.
- Suboptimal prediction of molecular properties: A biased dataset might overemphasize certain molecular properties (e.g., small molecules with known pharmacological properties) and ignore others (e.g., materials science compounds). As a result, the model may perform well in predicting specific chemical attributes (like drug-likeness) but poorly in others (like catalytic efficiency or solubility in non-standard solvents), limiting its usefulness across different domains of chemistry.
- Inaccurate reaction predictions: Biases in reaction datasets can cause CLMs to favor certain reaction mechanisms or conditions that are over-represented in the training data, leading to inaccurate or incomplete predictions for less common reactions. For example, if the dataset heavily features common organic reactions, the model may fail to predict outcomes in organometallic chemistry or biocatalysis.
- Challenges in handling rare or complex molecules: Dataset bias against rare, large, or structurally complex molecules can lead to poor handling of these types in real-world scenarios. This becomes a critical issue in areas like drug discovery or materials design, where unique molecular scaffolds or highly functionalized structures are often key to innovation.
- Overfitting and narrow focus: Models trained on biased data may overfit to specific molecular patterns or chemical properties, narrowing their focus. This overfitting makes the model less flexible and more prone to errors when faced with new tasks that require broader chemical knowledge or cross-domain adaptability.

Nevertheless, representative datasets for real-world challenges are of utmost importance for various reasons, such as:

- Capturing chemical diversity: Real-world chemical challenges involve a wide range of molecular structures, chemical reactions, and environmental conditions. Ensuring that datasets capture this diversity—spanning small molecules, macromolecules, organometallic compounds, and biologically relevant structures—allows the CLM to better predict real-world chemical behavior and outcomes. Representative datasets help prevent models from becoming overly specialized and allow them to remain versatile across various domains.
- Balanced molecular property distribution: It's essential that datasets contain molecules with a variety of properties, such as polarity, solubility, reactivity, and toxicity. This balance ensures the model can perform well in diverse tasks, from drug design to materials science, and enhances the model's robustness when predicting different chemical properties.
- Addressing underrepresented domains: Representative datasets ensure that underexplored areas of chemistry, such as sustainable catalysis, green chemistry, and rare molecular scaffolds, are included. Through this, CLMs are better prepared to tackle cutting-edge research challenges and can contribute to areas where innovative solutions are needed, such as renewable energy materials or environmentally friendly chemical processes.
- Improving transferability across tasks: Models trained on biased datasets may have limited ability to transfer learned knowledge to new tasks or domains. By ensuring the dataset is comprehensive and diverse, the model can more easily adapt to novel challenges, making it more useful across different branches of chemistry, from pharmaceuticals to environmental chemistry.
- Preventing systematic errors: A representative dataset helps in reducing the likelihood of systematic errors. For example, if a dataset disproportionately represents reactions or molecular structures from academic literature, it may underperform in industrial contexts where reactions may follow different protocols. Broadening the dataset to include real-world industrial reactions ensures the model's robustness in a variety of applications.

To ensure that CLMs are robust, flexible, and capable of addressing real-world chemical challenges, it is crucial to mitigate dataset bias and ensure newly proposed datasets are representative of the full range of chemical diversity, environments, and molecular properties. This enhances the model's generalizability, reduces errors, and ensures it performs optimally across various domains in chemistry, from drug discovery to materials science and environmental chemistry.

Future research should also focus on optimizing these models for better data efficiency, optimizing computational requirements, and enhancing model interpretability. Developing methods to better handle rare or novel molecular structures will further expand the applicability of transformers in computational chemistry. Additionally, integrating these CLMs with other ML approaches, such as graph neural networks (GNNs) [63], CNNs [79], could further improve their capability to understand and predict complex molecular properties.

9. Conclusion

Transformer-based models have revolutionized computational chemistry, offering robust tools for analyzing and predicting molecular properties from SMILES representations. Their ability to learn complex relationships or dependencies, scale to large datasets, and leverage contextual information positions them as one of the powerful tools in drug discovery, molecular design, and property prediction. CLMs have had a transformative impact on cheminformatics by revolutionizing the way chemical data is analyzed and interpreted. They have significantly enhanced the ability to predict molecular properties, design novel compounds, and streamline drug discovery by leveraging vast chemical datasets in ways that were previously unattainable. CLMs have enabled more accurate and scalable representations of molecular structures, improving both efficiency and precision in cheminformatics tasks. Beyond cheminformatics, the broader implications of CLMs extend to related fields such as materials science, where they assist in the discovery of new materials, and synthetic biology, where they aid in the design of biological molecules. Their ability to process and model complex chemical and biological information is driving innovation across multiple scientific disciplines.

As the field continues to evolve, ongoing innovations and addressing current limitations will be crucial to further enhance the potential capabilities and applications of these CLMs in cheminformatics. In general, this literature review summarizes the evolution of molecular representation methods, focusing on the emergence of transformer and attention-based CLMs. By providing a thorough overview of their architectures, applications, and challenges, this review aims to guide researchers and practitioners in leveraging these advanced CLMs for enhanced molecular analysis and prediction.

For the potential future directions of this review, we plan to address various important issues in our forthcoming work, such as: Integration with experimental chemistry: Explore how transformer-based CLMs can be integrated with experimental workflows, such as reaction optimization or compound screening. Discuss the potential for AI to accelerate real-time experimental design and feedback. Expansion to Multimodal Models: Address the potential for combining chemical data with other modalities (e.g., spectroscopy, biological data) to create more holistic models that better represent complex chemical systems and drug discovery pipelines. Ethical and Regulatory Considerations: Emphasize the growing importance of addressing ethical issues, such as dual-use risks, data biases, and the need for regulations governing AI in chemistry. This future direction could drive more responsible development and deployment of CLMs. Scalability and Resource Efficiency: Discuss advances in model pruning, quantization, and distributed training to address computational constraints. Highlight how these techniques can make large-scale models more accessible and environmentally sustainable. Personalized Medicine and Chemical Design: Suggest how future CLMs could be tailored for applications like personalized drug discovery or the design of custom materials, where predictions are finely tuned for individual or specialized use cases.

CRediT authorship contribution statement

Medard Edmund Mswahili: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Young-Seob Jeong:** Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Funding

Following are results of a study on the “Leaders in INdustry-university Cooperation 3.0” Project, supported by the Ministry of Education and National Research Foundation of Korea. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2020R1A6A1A12047945).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The Data Analysis and Artificial Intelligence (DAI) Lab http://dai.chungbuk.ac.kr/.

Data availability

No additional data was used for the research described in the article.

References

[1] D. Weininger, Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.

[2] J. Arús-Pous, et al., Randomized smiles strings improve the quality of molecular generative models, J. Cheminform. 11 (2019) 1–13.

[3] F. Grisoni, Chemical language models for de novo drug design: challenges and opportunities, Curr. Opin. Struct. Biol. 79 (2023) 102527.

[4] A.V. Sadybekov, V. Katritch, Computational approaches streamlining drug discovery, Nature 616 (2023) 673–685.

[5] J.J. Irwin, et al., Zinc20—a free ultralarge-scale chemical database for ligand discovery, J. Chem. Inf. Model. 60 (2020) 6065–6073.

[6] L. Yu, Y. Su, Y. Liu, X. Zeng, Review of unsupervised pretraining strategies for molecules representation, Brief. Funct. Genomics 20 (2021) 323–332.

[7] J. Wang, H. Wang, X. Wang, H. Chang, Predicting drug-target interactions via fm-dnn learning, Curr. Bioinform. 15 (2020) 68–76.

[8] X. Zeng, et al., Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest, Bioinformatics 36 (2020) 2805–2812.

[9] W.P. Walters, R. Barzilay, Applications of deep learning in molecule generation and molecular property prediction, Acc. Chem. Res. 54 (2020) 263–270.

[10] Z. Liu, et al., Ai-based language models powering drug discovery and development, Drug Discov. Today 26 (2021) 2593–2607.

[11] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[12] S. Wang, Y. Guo, Y. Wang, H. Sun, J. Huang, Smiles-bert: large scale unsupervised pre-training for molecular property prediction, in: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019, pp. 429–436.

[13] S. Chithrananda, G. Grand, B. Ramsundar, Chemberta: large-scale self-supervised pretraining for molecular property prediction, preprint, arXiv:2010.09885, 2020.

[14] B. Fabian, et al., Molecular representation learning with language models and domain-relevant auxiliary tasks, preprint, arXiv:2011.13230, 2020.

[15] K. Yang, et al., Analyzing learned molecular representations for property prediction, J. Chem. Inf. Model. 59 (2019) 3370–3388.

[16] A. Mayr, et al., Large-scale comparison of machine learning methods for drug target prediction on chembl, Chem. Sci. 9 (2018) 5441–5451.

[17] A.M. Dai, Q.V. Le, Semi-supervised sequence learning, Adv. Neural Inf. Process. Syst. 28 (2015).

[18] Z. Xu, S. Wang, F. Zhu, J. Huang, Seq2seq fingerprint: an unsupervised deep molecular embedding for drug discovery, in: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2017, pp. 285–294.

[19] X. Zhang, et al., Seq3seq fingerprint: towards end-to-end semi-supervised deep drug discovery, in: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018, pp. 404–413.

[20] A. Sherstinsky, Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network, Phys. D: Nonlinear Phenom. 404 (2020) 132306.

[21] R. Dey, F.M. Salem, Gate-variants of gated recurrent unit (gru) neural networks, in: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), IEEE, 2017, pp. 1597–1600.

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, preprint, arXiv:1810.04805, 2018.

[23] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training. OpenAI, 2018.

[24] T. Brown, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.

[25] S. Raschka, Y.H. Liu, V. Mirjalili, D. Dzhulgakov, Machine Learning with PyTorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python, Packt Publishing Ltd, 2022.

[26] A. Abaskohi, Navigating transformers: a comprehensive exploration of encoder-only and decoder-only models, right shift, and beyond (2023) Accessed: 2024-09-26.

[27] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Trans. Neural Netw. Learn. Syst. 33 (2021) 6999–7019.

[28] I. Lee, H. Nam, Infusing linguistic knowledge of smiles into chemical language models, preprint, arXiv:2205.00084, 2022.

[29] T. Guo, et al., What can large language models do in chemistry? A comprehensive benchmark on eight tasks, Adv. Neural Inf. Process. Syst. 36 (2024).

[30] J. Mao, et al., Transformer-based molecular generative model for antiviral drug design, J. Chem. Inf. Model. (2023).

[31] S. Adilov, Generative pre-training from molecules, ChemRxiv. (2021).

[32] R. Irwin, S. Dimitriadis, J. He, E.J. Bjerrum, Chemformer: a pre-trained transformer for computational chemistry, Mach. Learn.: Sci. Technol. 3 (2022) 015022.

[33] Z. Liu, et al., Molxpt: wrapping molecules with text for generative pre-training, preprint, arXiv:2305.10688, 2023.

[34] D. Oniani, et al., Emerging opportunities of using large language models for translation between drug molecules and indications, preprint, arXiv:2402.09588, 2024.

[35] R. Luo, et al., Biogpt: generative pre-trained transformer for biomedical text generation and mining, Brief. Bioinform. 23 (2022) bbac409.

[36] Y. Omote, K. Matsushita, T. Iwakura, A. Tamura, T. Ninomiya, Transformer-based approach for predicting chemical compound structures, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 154–162.

[37] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, B. Ramsundar, Chemberta-2: towards chemical foundation models, preprint, arXiv:2209.01712, 2022.

[38] S. Haroon, C. Hafsath, A. Jereesh, Generative pre-trained transformer (gpt) based model with relative attention for de novo drug design, Comput. Biol. Chem. 106 (2023) 107911.

[39] Xue, D. et al., X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis. bioRxiv 2020–12 (2020).

[40] S. Balaji, R. Magar, Y. Jadhav, et al., Gpt-molberta: Gpt molecular features language model for molecular property prediction, preprint, arXiv:2310.03030, 2023.

[41] Y. Fang, et al., Mol-instructions: a large-scale biomolecular instruction dataset for large language models, preprint, arXiv:2306.08018, 2023.

[42] S. Honda, S. Shi, H.R.Ueda, Smiles transformer: pre-trained molecular fingerprint for low data drug discovery, preprint, arXiv:1911.04738, 2019.

[43] T. Tran, C. Ekenna, Molecular descriptors property prediction using transformer-based approach, Int. J. Mol. Sci. 24 (2023) 11948.

[44] P. Morris, R. St. Clair, W.E. Hahn, E. Barenholtz, Predicting binding from screening assays with transformer network embeddings, J. Chem. Inf. Model. 60 (2020) 4191–4199.

[45] J. Ross, et al., Large-scale chemical language representations capture molecular structure and properties, Nat. Mach. Intell. 4 (2022) 1256–1264.

[46] L. Krasnov, I. Khokhlov, M.V. Fedorov, S. Sosnin, Transformer-based artificial neural networks for the conversion between chemical notations, Sci. Rep. 11 (2021) 14798.

[47] A. Yüksel, E. Ulusoy, A. Ünlü, T. Doğan, Selformer: molecular representation learning via selfies language models, Mach. Learn.: Sci. Technol. (2023).

[48] J. Born, M. Manica, Regression transformer enables concurrent sequence regression and generation for molecular language modelling, Nat. Mach. Intell. 5 (2023) 432–444.

[49] J. Li, X. Jiang, Mol-bert: an effective molecular representation with bert for molecular property prediction, Wirel. Commun. Mob. Comput. 2021 (2021) 1–7.

[50] Y. Liu, et al., Molrope-bert: an enhanced molecular representation with rotary position embedding for molecular property prediction, J. Mol. Graph. Model. 118 (2023) 108344.

[51] D. Chen, J. Zheng, G.-W. Wei, F. Pan, Extracting predictive representations from hundreds of millions of molecules, J. Phys. Chem. Lett. 12 (2021) 10793–10801.

[52] Y. Liu, et al., Roberta: a robustly optimized bert pretraining approach, preprint, arXiv:1907.11692, 2019.

[53] J. Lee, et al., Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[54] Y. Gu, et al., Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthcare 3 (2021) 1–23.

[55] N. O'Boyle, A. Dalke, Deepsmiles: an adaptation of smiles for use in machine-learning of chemical structures, ChemRxiv. (2018).

[56] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, R.G. Coleman, Zinc: a free tool to discover chemistry for biology, J. Chem. Inf. Model. 52 (2012) 1757–1768.

[57] S. Kim, et al., Pubchem 2019 update: improved access to chemical data, Nucleic Acids Res. 47 (2019) D1102–D1109.

[58] S. Kim, et al., Pubchem 2023 update, Nucleic Acids Res. 51 (2023) D1373–D1380.

[59] A. Gaulton, et al., The chembl database in 2017, Nucleic Acids Res. 45 (2017) D945–D954.

[60] A. Gaulton, et al., Chembl: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) D1100–D1107.

[61] Z. Wu, et al., Moleculenet: a benchmark for molecular machine learning, Chem. Sci. 9 (2018) 513–530.

[62] M.E. Mswahili, et al., Cocrystal prediction using machine learning models and descriptors, Appl. Sci. 11 (2021) 1323.

[63] M.E. Mswahili, G.E. Ndomba, K. Jo, Y.-S. Jeong, Graph neural network and bert model for antimalarial drug predictions using plasmodium potential targets, Appl. Sci. 14 (2024) 1472.

[64] S. Kim, S. Bae, Y. Piao, K. Jo, Graph convolutional network for drug response prediction using gene expression data, Mathematics 9 (2021) 772.

[65] U.V. Ucak, I. Ashyrmamatov, J. Lee, Improving the quality of chemical language model outcomes with atom-in-smiles tokenization, J. Cheminform. 15 (2023) 55.

[66] M. Domingo, M. Garcıa-Martınez, A. Helle, F. Casacuberta, M. Herranz, How much does tokenization affect neural machine translation?, preprint, arXiv:1812.08621, 2018.

[67] X. Li, D. Fourches, Smiles pair encoding: a data-driven substructure tokenization algorithm for deep learning, J. Chem. Inf. Model. 61 (2021) 1560–1569.

[68] J. Su, et al., Roformer: enhanced transformer with rotary position embedding, Neurocomputing 568 (2024) 127063.

[69] I. Goodfellow, et al., Generative adversarial networks, Commun. ACM 63 (2020) 139–144.

[70] C. Doersch, Tutorial on variational autoencoders, preprint, arXiv:1606.05908, 2016.

[71] M. Krenn, et al., Selfies and the future of molecular string representations, Patterns 3 (2022).

[72] D.S. Wigh, J.M. Goodman, A.A. Lapkin, A review of molecular representation in the age of machine learning, Wiley Interdiscip. Rev. Comput. Mol. Sci. 12 (2022) e1603.

[73] Z. Li, M. Jiang, S. Wang, S. Zhang, Deep learning methods for molecular representation and property prediction, Drug Discov. Today 27 (2022) 103373.

[74] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, Self-referencing embedded strings (selfies): a 100% robust molecular string representation, Mach. Learn.: Sci. Technol. 1 (2020) 045024.

[75] H. Abdel-Aty, I.R. Gould, Large-scale distributed training of transformers for chemical fingerprinting, J. Chem. Inf. Model. 62 (2022) 4852–4862.

[76] W. Gao, C.W. Coley, The synthesizability of molecules proposed by generative models, J. Chem. Inf. Model. 60 (2020) 5714–5723.

[77] K. Roy, In silico drug design: repurposing techniques and methodologies, Academic Press, 2019.

[78] E. Harigua-Souiai, et al., Deep learning algorithms achieved satisfactory predictions when trained on a novel collection of anticoronavirus molecules, Front. Genet. 12 (2021) 744170.

[79] M. Huang, et al., Nonlinear modeling of temperature-induced bearing displacement of long-span single-pier rigid frame bridge based on dcnn-lstm, Case Stud. Therm. Eng. 53 (2024) 103897.