Journal of Cheminformatics

RESEARCH

Open Access

Positional embeddings and zero-shot learning using BERT for molecular-property prediction

Medard Edmund Mswahili[1†], JunHa Hwang[1†], Jagath C. Rajapakse[2], Kyuri Jo[1*] and Young-Seob Jeong[1*]

Abstract

Recently, advancements in cheminformatics such as representation learning for chemical structures, deep learning (DL) for property prediction, data-driven discovery, and optimization of chemical data handling, have led to increased demands for handling chemical simplified molecular input line entry system (SMILES) data, particularly in text analysis tasks. These advancements have driven the need to optimize components like positional encoding and positional embeddings (PEs) in transformer model to better capture the sequential and contextual information embedded in molecular representations. SMILES data represent complex relationships among atoms or elements, rendering them critical for various learning tasks within the field of cheminformatics. This study addresses the critical challenge of encoding complex relationships among atoms in SMILES strings to explore various PEs within the transformer-based framework to increase the accuracy and generalization of molecular property predictions. The success of transformer-based models, such as the bidirectional encoder representations from transformer (BERT) models, in natural language processing tasks has sparked growing interest from the domain of cheminformatics. However, the performance of these models during pretraining and fine-tuning is significantly influenced by positional information such as PEs, which help in understanding the intricate relationships within sequences. Integrating position information within transformer architectures has emerged as a promising approach. This encoding mechanism provides essential supervision for modeling dependencies among elements situated at different positions within a given sequence. In this study, we first conduct pretraining experiments using various PEs to explore diverse methodologies for incorporating positional information into the BERT model for chemical text analysis using SMILES strings. Next, for each PE, we fine-tune the best-performing BERT (masked language modeling) model on downstream tasks for molecular-property prediction. Here, we use two molecular representations, SMILES and DeepSMILES, to comprehensively assess the potential and limitations of the PEs in zero-shot learning analysis, demonstrating the model's proficiency in predicting properties of unseen molecular representations in the context of newly proposed and existing datasets.

Scientific contribution

This study explores the unexplored potential of PEs using BERT model for molecular property prediction. The study involved pretraining and fine-tuning the BERT model on various datasets related to COVID-19, bioassay data,

[†]Medard Edmund Mswahili and JunHa Hwang contributed equally to this study.

*Correspondence:
Kyuri Jo
kyurijo@chungbuk.ac.kr
Young-Seob Jeong
ysjay@chungbuk.ac.kr
Full list of author information is available at the end of the article

and other molecular and biological properties using SMILES and DeepSMILES representations. The study details the pretraining architecture, fine-tuning datasets, and the performance of the BERT model with different PEs. It also explores zero-shot learning analysis and the model's performance on various classification and regression tasks. In this study, newly proposed datasets from different domains were introduced during fine-tuning in addition to the existing and commonly used datasets. The study highlights the robustness of the BERT model in predicting chemical properties and its potential applications in cheminformatics and bioinformatics.

Keywords Transformers, BERT, Positional embedding/encoding, Zero-shot learning, Molecular-property prediction, SMILES, DeepSMILES

Introduction

Predicting physicochemical molecular properties is a critical task in computational chemistry, driving research advancements in the domains of drug discovery, materials science, and chemical engineering [1–3]. Traditionally, these property predictions have relied predominantly on extensive experimental data and complex simulations, both time-consuming and resource-intensive approaches [4, 5]. However, with the advent of machine learning (ML) and deep learning techniques [6], these challenges have been overcome. In particular, leveraging the extensive available data, predictive models exhibiting excellent generalization capabilities across various chemical compounds have been developed using simplified molecular input line entry system (SMILES) strings in the field of cheminformatics [7, 8]. Consequently, data-driven ML approaches have recently gained traction for various tasks including chemical- and molecular-property prediction [9, 10]. Notably, leveraging the massive amounts of unlabeled SMILES data [11] and limited labeled SMILES data, ML techniques, particularly language models (LMs) [12–15], have been trained to learn insights or informative molecular representations from SMILES strings [16]. This has been accomplished by adopting a pretraining and fine-tuning framework or a semi-supervised learning architecture [17]. In this efficient framework, unsupervised pretraining using unlabeled SMILES data (molecules) is followed by fine-tuning using labeled data for specific downstream tasks. Given that obtaining abundant labeled molecular-property data through screening experiments and data labeling is both labor-intensive and resource-demanding, this approach provides a practical solution for the drug discovery pipeline [14, 18, 19].

In the field of natural language processing (NLP), various LM approaches have been extensively adopted, yielding promising prediction results in cheminformatics, particularly for accelerating and improving various molecular-property-prediction tasks [18, 20, 21]. Importantly, in recent years, transformer-based LMs, particularly the bidirectional encoder representations from transformer (BERT) models, have demonstrated exceptional performance in various NLP tasks in computational chemistry by effectively capturing contextual relationships within SMILES data [1, 2, 22–24]. These models particularly leverage generalized knowledge obtained during pretraining on extensive chemical datasets, such as SMILES representations, and are then fine-tuned on specific tasks. This approach has been shown to improve predictive performance for different tasks such as molecular properties or biological properties of chemical compounds [25–27]. Remarkably, this approach significantly enhances the generalization ability of the model, reducing the reliance on extensive, annotated chemical datasets. Notably, BERT's architecture [28], characterized by its self-attention mechanisms combined with masked language modeling (MLM) and next sentence prediction (NSP), allows it to effectively model dependencies between words or tokens in a given sequence. However, in the context of SMILES or DeepSMILES input representations, the NSP approach is often neglected [22]. A key innovation of BERT is its use of PEs, which encode the order of input tokens, thus enabling the model to distinguish between various sequences and comprehend the relative positioning of tokens.

Positional encoding and PEs are essential for representing sequential data, ensuring that models capture not only the presence of tokens but also their order [29]. Recent studies have underscored the effectiveness of PEs within transformer architectures, demonstrating their ability to accurately model the dependencies between elements or tokens across various orders or positions within a given sequence [29]. This concept extends beyond NLP to other domains where the order and structure of input data are of paramount importance, such as the representation of chemical compounds using SMILES and DeepSMILES [30]. Notably, PEs effectively encode the positional information contained in SMILES sequences, thereby improving the ability of the BERT pretrained model to identify and interpret potential information regarding molecular substructures for molecular-property prediction. By encoding the positional information of atoms and bonds in chemical representations such as SMILES

or DeepSMILES, transformer LMs such as the BERT model can potentially learn the underlying patterns governing chemical properties.

While BERT-based chemical LMs have demonstrated better scaling when applied to tremendously large unlabeled datasets and have shown promising performance across a wide range of chemical and cheminformatics tasks [31], the application and potential of PEs in this context remain underexplored, both in terms of the diversity of tasks they can handle and their domain of application such as zero-shot learning. Zero-shot learning-an approach wherein a model engages in predictions in classes or tasks not explicitly encountered during training. This is achieved using pre-existing knowledge and semantic relationships to generalize from known to unknown classes or tasks. For example, a model trained on text data might employ linguistic patterns and contextual understanding to analyze and categorize new chemical properties without having seen specific examples of those properties before. Zero-shot learning represents a promising direction for chemical-property prediction [32].

Although no single study has specifically focused on PEs and zero-shot learning using BERT for chemical-property prediction, several related studies provide relevant insights. For instance, Li et al. introduced MolBERT [1], which utilizes absolute PEs, while Liu et al. expanded this frame and developed MolRoPE-BERT [2], incorporating rotary PEs. Notably, both studies demonstrated notable improvements in chemical-property prediction, underscoring the significance of PEs. However, they share some common limitations, including inadequate evaluation of downstream and zero-shot learning tasks, limited pretraining datasets, and increased reliance on standard fine-tuning datasets. Future research must overcome these shortcomings by exploring more diverse and extensive pretraining datasets and conducting thorough evaluations of PEs across various tasks to comprehensively understand their potential and limitations.

In this study, to understand the potential of PEs, we adapt the BERT model with various PEs for chemical-property prediction across multiple real-world datasets. Specifically, we investigate ways in which the BERT architecture, traditionally used in NLP, can be effectively adapted to encode chemical information through a comprehensive experimental review of various PEs. Furthermore, we assess the potential of zero-shot learning for predicting the properties of novel compounds, offering insights into the generalization ability of the model across diverse chemical structures. Our findings suggest that transformer-based models equipped with appropriate embeddings and training strategies can serve as powerful tools in computational chemistry, providing efficient and accurate predictions of chemical properties. In summary, this study makes the following contributions.

- The study adopts a two-stage approach. First, the BERT model is pretrained on each types of position encoding and PE using SMILES representations. Second, the model is fine-tuned on various downstream tasks using both SMILES and DeepSMILES molecular representations. Furthermore, their performance on these tasks is compared to evaluate the potential of the BERT model and PEs in the context of zero-shot learning.
- In this study through a detailed experimental analysis, we compare different position encoding and PEs, including absolute, relative_key, relative_key_query, and sinusoidal PE, used in transformer-based models such as BERT for prediction of physicochemical-properties, and biological properties or biological activity.
- We introduce new datasets for fine-tuning from new domains, which are not typically included in standard fine-tuning databases [7] such as COVID-19, antimalarial drugs, and cocrystal formation using active pharmaceutical ingredients (APIs) and co-formers, are incorporated.

Materials and methods

Dataset

The BERT-based model was pretrained and fine-tuned using various datasets to evaluate its performance across cheminformatics and bioinformatics tasks. The datasets used for pretraining and fine-tuning are summarized in Table 1. Fine-tuning datasets were selected to address either regression or classification tasks, focusing on molecular properties, biological activities and drug responses. The datasets include SMILES and DeepSMILES [30] representations of chemical compounds and are categorized into two groups based on their use in pretraining and fine-tuning strategies, as shown in Fig. 1. This categorization supports the effective utilization and application of the BERT-based LM. Pretraining datasets play a critical role in assisting the model learn patterns and influence its performance and generalization to unseen data. The effectiveness of the pretraining dataset relies on meticulous data collection, pre-processing, and cleaning to ensure that the model learns meaningful patterns. In this study, the dataset used for model pretraining was sourced from publicly available databases and previous studies, as indicated in Table 1. As depicted, this dataset comprises 7,949,003 SMILES-string instances, following the removal of 411,621 duplicated instances.

Table 1 Dataset used for pretraining and fine-tuning of the BERT-based model

Pretraining data	Fine-tuning data	Task target	Task type	# Task	# Compounds
ZINC [33, 34]	Physical Chemistry [7]	ESOL	Regression	1	1128
PubChem [35, 36]		FreeSolv	Regression	1	642
ChEMBL [37, 38]		Lipophilicity	Regression	1	4200
Research studies [1, 2]	Physiology [7]	BBBP	Classification	1	2039
		Tox21	Classification	12	7831
		ClinTox	Classification	2	1478
		SIDER	Classification	27	1427
	Newly proposed	Antimalarial [39]	Classification	1	4794
		Cocrystals [40]	Classification	1	3282
		COVID [41]	Classification	1	740
		COVID-19 [42]	Classification	1	2601

Bold values highlight the newly proposed datasets introduced in this study, representing the key focus areas

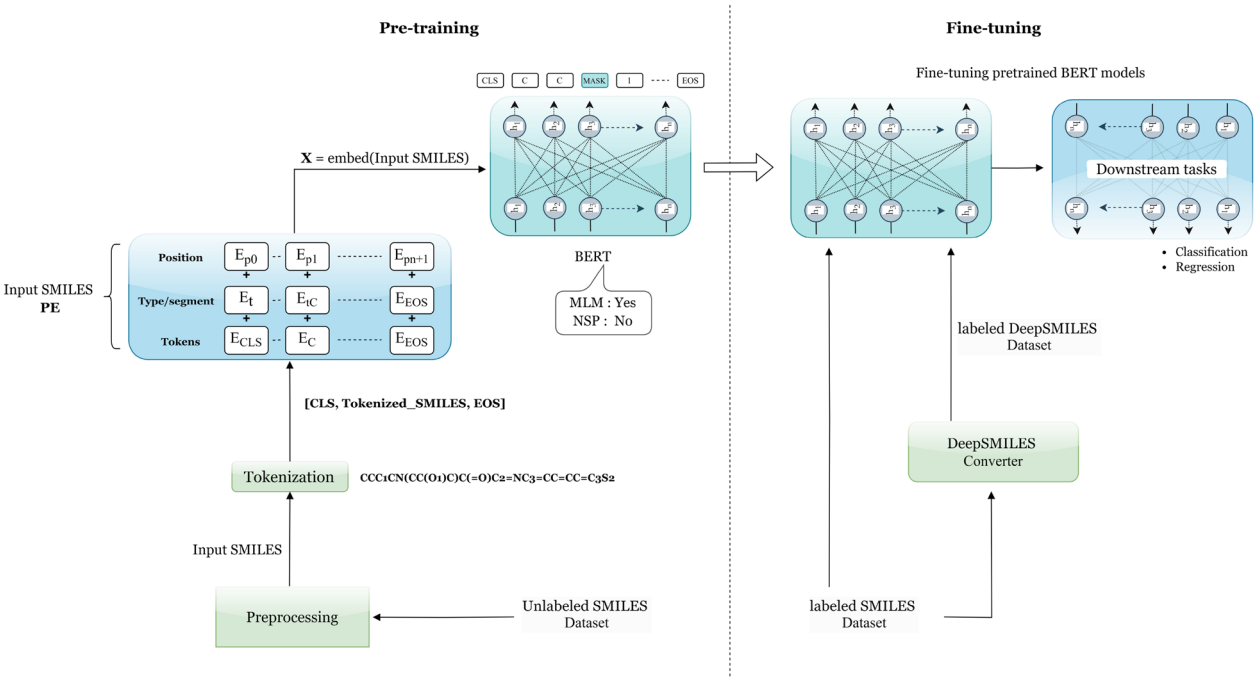


Fig. 1 Pretraining and fine-tuning architectures employed in BERT model for SMILES and DeepSMILES (zero-shot learning analysis). During the fine-tuning phase, we utilized both SMILES and DeepSMILES representations by employing a Python module (converter) to transform well-formed SMILES strings into DeepSMILES format [30]

The fine-tuning phase employed datasets tailored for various tasks, such as molecular-property prediction and drug discovery, in cheminformatics and bioinformatics. Fine-tuning datasets are important for adapting a pre-trained model to specialized tasks or domains, allowing the model to learn task-specific patterns and improve its performance. In the fine-tuning datasets, SMILES and DeepSMILES strings are employed to represent the input molecules, as shown in Fig. 1. A detailed breakdown of the fine-tuning datasets utilized in this study, along with their corresponding tasks, is outlined in Table 1. In addition to the commonly used fine-tuning datasets, this study adopted several new datasets (i.e., Newly proposed) to evaluate the model's performance on more specific tasks.

Antimalarial [39]

The Antimalarial dataset comprises experimentally verified antimalarial drug candidates sourced from public chemical databases. It includes compounds labeled

"active", indicating successful reactivity against the parasite species *Plasmodium falciparum*, and compounds labeled "inactive", that have no reactivity against the parasite species *Plasmodium falciparum*.

Cocrystals [40, 43]

The Cocrystal dataset, wherein each API can chemically interact with multiple coformers, includes 79 unique APIs and 462 unique coformers, resulting in 1641 instances. In this dataset, 880 interactions are labeled "0", indicating instances where no cocrystal formation occurred, while 761 interactions are labeled "1" to signify successful cocrystal formation. The raw data were collected from various sources, including previous research studies, experimental documentation, and ongoing experiments related to cocrystal formation. Notably, in the case of the cocrystals dataset, the input sequences processed by the inherited BERT model consisted of two distinct compounds: the API and coformers. Unlike other datasets that contained a single compound per sequence, the cocrystals dataset required concatenation of the SMILES or DeepSMILES representations of both compounds, resulting in significantly longer single sequences.

COVID [41]

The COVID dataset comprises 740 chemical compounds that have exhibited experimental activity against various coronavirus targets. These targets include SARS-CoV, MERS-CoV, SARS-CoV-2, ORF1ab/ORF1a polyproteins, ORF1ab polyprotein (Betacoronavirus England 1), surface glycoprotein (SARS-CoV-2), and replicase polyprotein 1ab (SARS-CoV-2). This dataset is curated from public databases of bioactive molecules with drug-like properties. We further expanded this dataset by including information regarding the structure of the viral 3-chymotrypsin-like cysteine protease ($3CL^{pro}$) enzyme, a proven drug target essential for coronavirus replication and life cycle of SARS-Cov-2. Notably, SARS-Cov-2 has a genome sequence similar to that of other members of the beta-coronavirus group such as SARS-Cov, MERS-CoV, and bat coronavirus.

COVID-19 [42]

The COVID-19 data collection process involved three main approaches:

- Literature mining: Data were gathered from peer-reviewed studies on molecules identified as anticoronavirus agents, focusing on two beta-coronavirus species: SARS-CoV (2003) and SARS-CoV-2 (2019).
- Cocrystal data retrieval: Information regarding molecules cocrystallized with SARS-CoV and SARS-CoV-2 proteins, such as 3CL-protease and papain-like protease, was obtained from the research collaboratory for structural bioinformatics (RCSB PDB). Activity data for these inhibitors were also sourced from relevant scientific publications.
- Bioassay data retrieval: Data were extracted from bioassays in the PubChem database, prioritizing those targeting SARS-CoV-2 or similar molecular targets, particularly large bioassays on other coronaviruses. This included viral growth inhibition and cell-based tests targeting specific viral enzymes.

The dataset comprised 1301 active molecules labeled "1", and 1300 inactive molecules labeled "0". The balanced distribution of active and inactive compounds ensured a robust evaluation of our model's performance, as it allowed for equitable comparison of predictions across both classes.

The remaining fine-tuning datasets utilized in this study were discussed by Wu et al. [7] as shown in Table 1. These new datasets were selected to facilitate a comprehensive assessment of the capabilities of the BERT-based model across various molecular and biological properties. The diverse nature of the datasets ensures that the model is tested on both regression and classification tasks, providing a thorough evaluation of its performance in cheminformatics and bioinformatics applications.

Class weights for imbalance datasets

Zhu et al. [44], class weights are typically computed to address the issue of class imbalance in datasets such as Side Effect Resource (SIDER), Blood–Brain Barrier Penetration (BBBP) or Tox21 (a toxicity dataset). These weights are used to adjust the importance of each class during training, ensuring that the minority class contributes proportionally to the loss function and the model learns its features effectively. In this study, we used common method for calculating class weights which is based on the inverse of the class frequencies. For example, if a dataset contains N samples and n_c samples for class c, the weight for that class is computed as shown in Eq. 1.

$$w_c = \frac{N}{(|C| \cdot n_c)}, \tag{1}$$

where $|C|$ is the total number of classes. This approach assigns higher weights to underrepresented classes, making their contribution to the loss function equal to that of more frequent classes.

Pretraining architecture

Our model leverages a BERT-base architecture specifically tailored for SMILES or DeepSMILES representations. This architecture includes several key

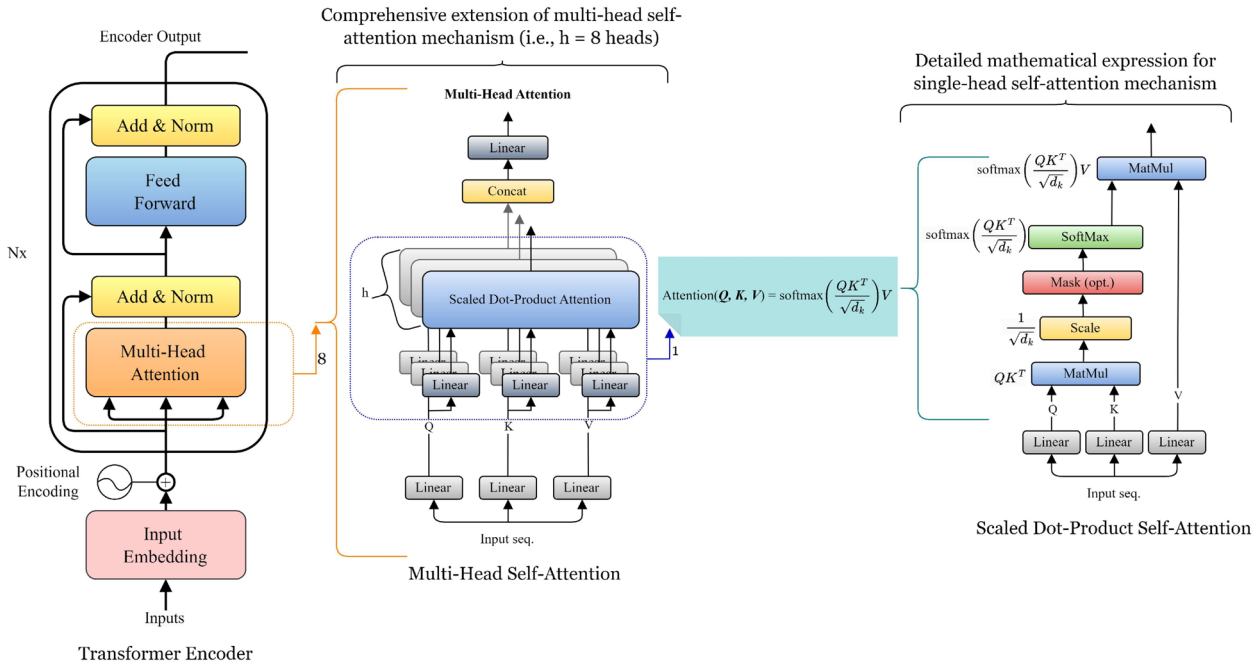


Fig. 2 PEs, position encoding, and Self-attention mechanism employed in BERT architecture (Position Encoding resembles both position embeddings and position encoding utilized in this study)

components such as an embedding layer to translate input tokens (SMILES or DeepSMILES strings) into continuous vector representations. In this context, various PEs, such as absolute, relative_key, and relative_key_query, and positional encodings such as sinusoidal positional encodings were investigated.

The transformer block or encoder layer-comprising multiple layers of self-attention mechanisms, also known as multi head self-attention and feed-forward neural networks-constituted the core of our model, as illustrated in Fig. 2. Notably, this structure enables the model to capture the intricate dependencies and relationships within SMILES or DeepSMILES sequences. Finally, the MLM task, a prediction head is used to reconstruct masked tokens based on their context.

Tokenizer

We adopted the SmilesTokenizer module from DeepChem [45, 46] to generate embedding features (**X**), as depicted in Fig. 1. The SmilesTokenizer module is derived from the implementation of the BertTokenizer class in Hugging Face's Transformers library. This tokenizer predominantly employs the byte-pair encoding (BPE) tokenization strategy from the Hugging Face tokenizers library [23, 47]. In particular, it executes a WordPiece tokenization algorithm on the collected SMILES strings, utilizing the tokenization SMILES regex developed by Schwaller et al. [48, 49].

PEs and positional encoding

Positional encoding and PEs are essential for representing sequential data and understanding intricate relationships within sequences, ensuring that models capture not only the presence of tokens but also their order [29].

Absolute PE [50, 51]

These embeddings represent the absolute positions of tokens in a sequence. For example: If a SMILES or DeepSMILES token is the 3rd word in a SMILES or DeepSMILES sequence, its position embedding specifically encodes the number 3. The absolute position embedding in self-attention [52] is as shown in Eq. 2.

$$x_i = t_i + s_i + w_i \tag{2}$$

where x_i, $i \in \{0, ..., n-i\}$ is the input embedding to the first transformer layer, t_i, s_i and $w_i \in \mathbb{R}^{d_x}$ are the token embeddings, segment embeddings, and absolute position embeddings respectively. However s_i is not incorporated in our case since we are only utilizing single SMILES or DeepSMILES sequence, and t_i and w_i are learnable parameters.

Relative key PE [50, 51]

This technique encodes the relative positions or distances between tokens into the self-attention mechanism. It often incorporates sinusoidal functions or learnable parameters. For example, If token A in this case SMILES or DeepSMILES is two positions away from token B in SMILES or DeepSMILES sequences, the model captures this relative distance (e.g., +2 or −2). A notable implementation was introduced by Shaw et al. [53], which adds position-specific information into the attention computation via Eq. 3:

$$e_{ij} = q_i^T k_j + a_{ij}, \tag{3}$$

where e_{ij} attention score, q_i and k_j are the query and key vectors, and a_{ij} represents the relative position embedding based on the distance between tokens i and j. This method allows the model to focus on how tokens relate to each other based on their relative distances.

Relative key-query PE [29, 50]

Building on the concept of relative position embeddings, this approach models interactions not just between tokens and their relative distances but also between keys, queries, and relative position embeddings simultaneously. This approach combines relative positions of both the keys and the queries in the attention mechanism as shown in Eq. 4. For example: If token A (as a query) in a SMILES or DeepSMILES sequence attends to token B (as a key) and token B is three positions ahead, this specific relative position is encoded.

$$q_i' = R(q_i, p_i), \quad k_j' = R(k_j, p_j), \tag{4}$$

where R is a rotational matrix applied to query and key vectors based on their positional indices. Since this approach captures both query and key position information, it enhances the model's ability to capture relative positional dependencies in a computationally efficient manner and assists in improving the model's contextual understanding.

Sinusoidal position encoding [52]

This technique is used in transformer models to provide information about the order of input tokens, since transformers lack inherent sequence awareness. It involves assigning each position in a sequence a unique vector, where each element of the vector is computed using sine or cosine functions of different frequencies. Specifically, the even-indexed dimensions use sine functions, and the odd-indexed dimensions use cosine functions as in shown in Eq. 5

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right), \tag{5}$$

where pos is the position index and d is the embedding dimension. This method provides a fixed encoding that is consistent across different input sequences. This encoding ensures that the position representations are continuous and periodic, allowing the model to capture relative distances between tokens and generalize to sequences longer than those it was trained on.

Zero-shot learning analysis

Zero-shot learning, which involves making predictions for classes or tasks not encountered during training, offers a promising approach for predicting chemical properties [32, 54]. In this study, we explored zero-shot learning by fine-tuning models initially trained using SMILES data. We utilized initial SMILES data from the collected fine-tuning dataset and converted them into DeepSMILES data for zero-shot learning analysis using the DeepSMILES converter [30]. To comprehensively evaluate the capabilities and potential of our models, we fine-tuned our best-performing models using both SMILES and DeepSMILES datasets, as illustrated in Fig. 1. Notably, DeepSMILES is an extension of SMILES designed to be more compatible with deep learning models. This framework simplifies some aspects of SMILES to facilitate neural network processing [30]. This approach allows us to assess the adaptability and generalization ability of our models to novel, related datasets, offering insights into their zero-shot learning capabilities.

Experiments and results

Experiments on BERT pretraining

In this study, standard BERT pretraining and fine-tuning procedures were adopted to train our model, specifically BERT for each PE. Notably, our implementation of the BERT model involves 12 attention heads and 12 layers, resulting in 144 distinct attention mechanisms. Notably, during the pretraining phase without NSP, the model is pretrained using only MLM on an extensive corpus of pre-processed SMILES data obtained from various sources, as outlined in Table 1. In this case, MLM masks 15% of the tokens in each input string, and the objective is to enable the BERT model to predict masked tokens from the input sequences. We used a maximum sequence length and vocabulary size of 512 and 592 tokens, respectively. To facilitate large-scale pretraining, we pre-processed and shuffled the remaining unique SMILES strings after removing duplicates from our dataset. The BERT model for each PE was trained over five epochs to avoid overfitting, and the best-performing model was utilized for subsequent fine-tuning on downstream tasks.

Table 2 Optimal parameters for pretaining and fine-tuning of the BERT model on SMILES and DeepSMILES data

Parameters	Pretraining	Fine-tuning	Position encoding/PEs
Learning rate	1e−4	5e−6	
Batch size	16	16	
Warm-up ratio	0.016	0.1	
Weight decay	0.01	0.01	
Number of epochs	5	10	
Optimizer	AdamW	AdamW	
Warm up schedular	Linear	Linear	
Number of parameters	85,054,464	86,496,002	Absolute
	85,840,128	87,281,666	Relative_key
	85,840,128	87,281,666	Relative_key_query
	85,054,464	86,496,002	Sinusoidal [52]

Table 3 Comparison results of different PEs and position encoding methods during pretraining of the BERT model on SMILES data

PEs and position encoding	Training time (h)	Optimal learning rate	Accuracy
Absolute	167	8.14e−5	0.9568
Relative_key	180	8.68e−6	0.9746
Relative_key_query	120	4.59e−6	**0.9763**
Sinusoidal [52]	105	1.67e−7	0.9755

Bold value denotes the best-achieved performance for clarity and emphasis

Fine-tuning of the architecture

Fine-tuning experiments were conducted on classification and regression tasks using the best-performing pretrained models. Notably, these pretrained models were fine-tuned according to the parameter settings detailed in Table 2 on specific datasets. These datasets covered various tasks, including the prediction of chemical properties, bioactivity of chemical compounds against varying targets, and drug discovery, as outlined in Table 1. Each fine-tuning dataset was split into training, validation, and test sets with proportions of 80%, 10%, and 10%, respectively. The performance metrics utilized during fine-tuning involved the test loss, accuracy, precision, recall, and F1-score for classification tasks, as well as test loss and root mean squared error (RMSE) for regression tasks. The pretrained models were evaluated on several tasks using fine-tuning datasets from MoleculeNet [7]. However, most existing datasets struggle with complex tasks owing to data scarcity issues and highly imbalanced classification. To address this, we employed a balanced class-weighted function during fine-tuning to handle imbalanced classes during classification tasks. Furthermore, as indicated in Table 1, we proposed new balanced datasets to evaluate model performance on additional tasks such as predicting anti-COVID drugs for all variants (i.e., SARS-CoV-2, and MERS-CoV, and bat coronavirus), antimalarial drugs against *Plasmodium falciparum*, and cocrystal formation using API and co-formers.

Pre-training BERT on SMILES

Tables 2 and 3 detail training parameter settings, performance metrics, evaluation metrics, and training time for each PE. The pretraining results of the proposed BERT model for each PE are summarized in Table 3. All PEs integrated into the BERT model demonstrated high accuracy, with all exceeding 95%. The relative_key_query PE showed slightly higher performance compared to the relative_key PE and sinusoidal positional encoding, although the differences in accuracy across the PEs were generally small, as indicated in Table 3.

Fine-tuning BERT on SMILES vs DeepSMILES (Zero-shot)

The performance of the BERT model with various positional encoding and PE methods was evaluated and fine-tuned on multiple classification and regression tasks using both SMILES and DeepSMILES representations for zero-shot learning analysis. The key performance metrics in this case included test loss, accuracy, and F1-score for classification tasks and test loss and RMSE for regression tasks. The hyperparameters for our experiments were selected using empirical methods. We followed a trial-and-error approach to identify the most promising parameters. This process involved systematically adjusting and testing various hyperparameter values to optimize the performance of our models. By iterating through various configurations, we fine-tuned the parameters and obtain improved results. This empirical approach, albeit time-consuming, ensured that we explored a wide range of possibilities to determine the optimal settings for our specific experimental context.

Relative key query PE

The fine-tuning results of the BERT model corresponding to the relative_key_query PE in classification tasks are outlined in Table 4. Notably, the model achieved high accuracy across all tasks, demonstrating notable performance on datasets such as ClinTox and Tox21, indicating its robustness, as illustrated in Supplemental Figure S1. For instance, the model achieved similar accuracy (i.e., 0.9262 and 0.9394, respectively) and F1-score (i.e., 0.9617 and 0.9688, respectively) values on the ClinTox and Tox21 datasets, using both SMILES and DeepSMILES

Table 4 Performance of fine-tuned BERT with relative_key_query PE on various datasets

Task	Data	Sequence	Test loss	Accuracy	Precision	Recall	F1-score
Classification	Malaria	SMILES	0.4858	0.8017	0.8118	0.6866	**0.7439**
		DeepSMILES	0.5203	0.7495	0.7793	0.5622	0.6532
	COVID	SMILES	0.5851	0.7568	0.8484	0.6829	**0.7568**
		DeepSMILES	0.4855	0.7568	0.8485	0.6829	**0.7568**
	COVID-19	SMILES	0.4855	0.7885	0.7881	0.7561	**0.7718**
		DeepSMILES	0.5171	0.7462	0.7568	0.6829	0.7179
	Cocrystals	SMILES	0.6089	0.6463	0.6102	0.5070	0.5538
		DeepSMILES	0.6011	0.6402	0.5882	0.5634	**0.5755**
	BBBP^{c_w}	SMILES	0.5876	0.7171	0.8039	0.8146	0.8092
		DeepSMILES	0.5422	0.7756	0.8571	0.8344	**0.8456**
	BBBP	SMILES	0.4679	0.7512	0.7475	1.0000	**0.8555**
		DeepSMILES	0.5592	0.7366	0.7366	1.0000	0.8483
	ClinTox	SMILES	0.4511	0.9262	0.9262	1.0000	**0.9617**
		DeepSMILES	0.4561	0.9262	0.9262	1.0000	**0.9617**
	Tox21^{c_w}	SMILES	0.6579	0.9040	0.9426	0.9560	**0.9493**
		DeepSMILES	0.6685	0.8419	0.9357	0.8931	0.9139
	Tox21	SMILES	0.3477	0.9394	0.9394	1.0000	**0.9688**
		DeepSMILES	0.2884	0.9394	0.9394	1.0000	**0.9688**

Bold values denote the best-achieved performance for clarity and emphasis

c_w class-weighted function, *DeepSMILES* the zero-shot learning analysis of BERT

Table 5 Performance of the fine-tuned BERT using RMSE for all PEs/position encoding on regression tasks

Data	Sequence	Relative_key_query	Sinusoidal	Relative_key	Absolute
ESOL	SMILES	0.6185	**0.5883**	0.7878	0.5983
	DeepSMILES	0.6557	0.6256	0.8431	**0.5584**
FreeSolv	SMILES	**1.8858**	2.0491	2.6242	2.4169
	DeepSMILES	2.1103	2.1572	2.0209	**1.9840**
Lipophilicity	SMILES	**0.5704**	0.5732	0.5716	0.6025
	DeepSMILES	**0.6333**	0.6707	0.6857	0.6747

Bold values denote the best-achieved performance for clarity and emphasis

DeepSMILES zero-shot learning analysis of BERT, *RMSE* Root Mean Squared Error

representations. In the Tox21 dataset, with imbalanced class instances, a class-weight function set to a balanced ratio (denoted as Tox21^{c_w}) was adopted to evaluate performance outcomes. On newly proposed datasets (such as malaria, COVID, COVID-19, and cocrystals), BERT achieved similar and comparable performance with both SMILES and DeepSMILES representations. Although we hypothesized that DeepSMILES (i.e., for zero-shot learning analysis) would yield less robust results due to the model being pretrained only on SMILES representations, the differences in downstream performance between SMILES and DeepSMILES were minimal in most tasks.

This observation is detailed in Table 4 and illustrated in Supplemental Figure S1, highlighting comparable efficacy across both encoding methods.

Similarly, in regression tasks, the BERT model using the relative_key_query PE demonstrated robust performance in predicting chemical properties, as detailed in Table 5 and Supplemental Figure S2. The model achieved an RMSE of 0.5704 for the Lipophilicity dataset using SMILES representations and 0.6333 on zero-shot learning using DeepSMILES representations. The model also demonstrated reliable predictive power, with RMSE values of 0.6185 for the ESOL dataset using SMILES representations and 0.6557 using DeepSMILES for zero-shot learning. This demonstrated the capability of the model to handle diverse regression tasks effectively. However, for the FreeSolv dataset, as illustrated in Supplemental Figure S2, the model exhibited slightly poorer predictive

Table 6 Performance results of fine-tuned BERT with sinusoidal positional encoding on various datasets

Task	Data	Sequence	Test loss	Accuracy	Precision	Recall	F1-score
Classification	Malaria	SMILES	0.5255	0.7537	0.7456	0.6269	0.6811
		DeepSMILES	0.5199	0.7599	0.7590	0.6269	**0.6866**
	COVID	SMILES	0.6062	0.7568	0.7805	0.7805	**0.7805**
		DeepSMILES	0.5300	0.7703	0.8750	0.6829	0.7671
	COVID-19	SMILES	0.4529	0.8115	0.8190	0.7724	**0.7950**
		DeepSMILES	0.5056	0.7615	0.7607	0.7236	0.7417
	Cocrystals	SMILES	0.5664	0.7134	0.6667	0.6761	0.6713
		DeepSMILES	0.4838	0.7500	0.6875	0.7746	**0.7285**
	BBBP^{c_w}	SMILES	0.5596	0.8439	0.8742	0.9205	**0.8967**
		DeepSMILES	0.5382	0.7024	0.9245	0.6490	0.7626
	BBBP	SMILES	0.5280	0.7463	0.7438	1.0000	**0.8531**
		DeepSMILES	0.5285	0.7415	0.7500	0.9735	0.8473
	ClinTox^{c_w}	SMILES	0.6518	0.8859	0.9291	0.9493	**0.9391**
		DeepSMILES	0.6036	0.7718	0.9815	0.7681	0.8618
	ClinTox	SMILES	0.4534	0.9195	0.9315	0.9855	0.9577
		DeepSMILES	0.4210	0.9262	0.9262	1.0000	**0.9617**
	Tox21^{c_w}	SMILES	0.6502	0.8316	0.9530	0.8630	0.9058
		DeepSMILES	0.6466	0.9321	0.9402	0.9906	**0.9647**
	Tox21	SMILES	0.3963	0.9380	0.9380	1.0000	**0.9680**
		DeepSMILES	0.2207	0.9380	0.9380	1.0000	**0.9680**
	SIDER^{c_w}	SMILES	0.5005	0.7892	0.7687	0.9759	**0.8600**
		DeepSMILES	0.4697	0.5594	0.6708	0.6596	0.6651
	SIDER	SMILES	0.5908	0.7852	0.7755	0.9518	0.8546
		DeepSMILES	0.5054	0.7892	0.7687	0.9759	**0.8600**

Bold values denote the best-achieved performance for clarity and emphasis

c_w class-weighted function, *DeepSMILES* zero-shot learning analysis of BERT

performance compared to its performance in other regression tasks.

Sinusoidal positional encoding

The results of BERT using sinusoidal positional encoding in classification and regression tasks are summarized in Tables 5 and 6 respectively. Notably, this positional encoding demonstrated competitive performance, with notably high accuracy and F1-score values of 0.9380 and 0.9680, respectively, on the Tox21 dataset using the SMILES and DeepSMILES representations. Notably, in zero-shot learning scenarios on the ClinTox and SIDER datasets, using DeepSMILES yielded slightly better results compared to using SMILES, resulting in F1-score values of 0.9577 for SMILES and 0.9617 for DeepSMILES and 0.8546 for SMILES and 0.8600 for DeepSMILES, respectively, as depicted in Supplemental Figure S3. When employing a class-weight function on imbalanced datasets (such as BBBP^{c_w}, ClinTox^{c_w}, Tox21^{c_w}, and SIDER^{c_w}), DeepSMILES outperformed SMILES in zero-shot learning on the Tox21^{c_w} dataset with an F1-score of 0.9647 compared to an F1-score of 0.9058 for SMILES representations. Meanwhile, on the ClinTox^{c_w} dataset, the performance gap between SMILES and DeepSMILES representations in zero-shot learning was narrow, particularly compared to the BBBP^{c_w} and SIDER^{c_w} datasets, as indicated in Supplemental Figure S4.

In regression tasks, the performance of the model resembled that in the case with the relative_key_query PE. Overall, the model demonstrated robust performance in predicting chemical properties, as detailed in Table 5 and Supplemental Figure S5. The performance of the model in terms of the loss and RMSE was relatively consistent across the ESOL and Freesolv datasets. However, on the Lipophilicity dataset, the model achieved an RMSE of 0.5732 using SMILES and an RMSE of 0.6707 using DeepSMILES.

Relative key PE

As detailed in Table 7, the BERT model demonstrated optimal performance in terms of the accuracy and F1-score on the Tox21 classification task using both SMILES and zero-shot DeepSMILES representations. Specifically, on the ClinTox dataset, the model peak

Table 7 Performance of fine-tuned BERT with "relative_key" PE on various datasets

Task	Data	Sequence	Test loss	Accuracy	Precision	Recall	F1-score
Classification	Malaria	SMILES	0.4643	0.7787	0.7568	0.6965	**0.7254**
		DeepSMILES	0.5220	0.7557	0.7360	0.6517	0.6913
	COVID	SMILES	0.4617	0.7973	0.8824	0.7317	0.8000
		DeepSMILES	0.4425	0.8108	0.9091	0.7317	**0.8180**
	COVID-19	SMILES	0.4526	0.7962	0.7917	0.7724	**0.7819**
		DeepSMILES	0.4986	0.7808	0.7750	0.7561	0.7654
	Cocrystals	SMILES	0.5827	0.7134	0.6935	0.6056	0.6466
		DeepSMILES	0.5249	0.7012	0.6410	0.7042	**0.6711**
	BBBP^{c_w}	SMILES	0.5651	0.7707	0.8210	0.8808	0.8498
		DeepSMILES	0.5436	0.7902	0.8913	0.8146	**0.8512**
	BBBP	SMILES	0.2910	0.8585	0.8861	0.9272	0.9061
		DeepSMILES	0.2836	0.8780	0.8938	0.9470	**0.9196**
	ClinTox	SMILES	0.0804	0.9799	1.0000	0.7857	0.8800
		DeepSMILES	0.0421	0.9866	1.0000	0.8571	**0.9231**
	Tox21	SMILES	0.3964	0.9380	0.9393	0.9984	0.9680
		DeepSMILES	0.2282	0.9394	0.9394	1.0000	**0.9688**

Bold values denote the best-achieved performance for clarity and emphasis

c_w class-weighted function, *DeepSMILES* zero-shot learning analysis of BERT

accuracy scores of 0.9799 and 0.9866 using SMILES and DeepSMILES representations, respectively. Meanwhile, for the newly proposed datasets, using zero-shot DeepSMILES representations yielded better outcomes compared to SMILES representations in two classification tasks but demonstrated comparable performance in others tasks. Overall, the accuracy and F1-score of the method demonstrated a similar trend, with DeepSMILES generally yielding higher values for most tasks compared to SMILES, as depicted in Supplemental Figure S6 and Table 7.

As illustrated in Table 5 and Supplemental Figure S7, the model with relative_key PE demonstrated similar performance trends in terms of the loss and RMSE as when using the relative_key_query PE and sinusoidal positional encoding for regression tasks. Specifically, compared to the Lipophilicity dataset, the model demonstrated relatively similar performance in terms of the loss and RMSE on the ESOL and FreeSolv datasets when using SMILES and DeepSMILES representations. Specifically, for the Lipophilicity dataset, the model achieved an RMSE of 0.5716 using SMILES and 0.6857 on zero-shot learning using DeepSMILES. Furthermore, when using the relative_key PE, using DeepSMILES for zero-shot learning analysis yielded better outcomes compared to SMILES on the FreeSolv dataset, as depicted in Supplemental Figure S7 and Table 7. This indicates a slightly greater performance variability when applying zero-shot learning with DeepSMILES in this context.

Absolute PE

Tables 8 and 5 present the results of the BERT model using the absolute PE for classification and regression tasks respectively. Notably, in classification tasks, the proposed model achieved the highest performance accuracy (i.e., 0.9365) and F1-score (i.e., 0.9672) on the Tox21 dataset, followed by the ClinTox dataset. Furthermore, when a class-weight function was implemented on imbalanced datasets such as BBBP^{c_w} and Tox21^{c_w}, the differences in the accuracy and F1 scores of the model when using the SMILES and DeepSMILES representations in zero-shot analysis were minimal, resulting in similar loss values on both datasets. Unlike other PEs, when the model adopted the absolute PE, using DeepSMILES for zero-shot learning yielded better outcomes compared to SMILES on two of the regression tasks, as indicated in Table 5. In summary, the BERT exhibited slightly better performance on most regression tasks when using the absolute PE.

We evaluate the performance of BERT transformer-encoder-based models using zero-shot learning (DeepSMILES) and different positional encoding and PEs across various classification and regression tasks, identifying distinct performance patterns based on the choice of molecular representations. The Table 9 compares F1 scores for classification tasks and RMSE for regression tasks across different types of position encoding and PEs using BERT. The higher prediction errors on the FreeSolv dataset observed in our experiments shown in Table 9 and Fig. 3 could stem from several factors inherent to

Table 8 Performance of fine-tuned BERT with "absolute" on various datasets

Task	Data	Sequence	Test loss	Accuracy	Precision	Recall	F1 score
Classification	Malaria	SMILES	0.5205	0.7537	0.7677	0.5920	**0.6685**
		DeepSMILES	0.5505	0.7307	0.7368	0.5572	0.6346
	COVID	SMILES	0.6077	0.7703	0.8529	0.7703	**0.7733**
		DeepSMILES	0.4851	0.7703	0.8529	0.7073	**0.7733**
	COVID-19	SMILES	0.4163	0.8154	0.8049	0.8049	**0.8049**
		DeepSMILES	0.4903	0.7577	0.7500	0.7317	0.7407
	Cocrystals	SMILES	0.6089	0.6463	0.6102	0.5070	**0.5538**
		DeepSMILES	0.6203	0.6098	0.5574	0.4789	0.5152
	BBBP^{c_w}	SMILES	0.6037	0.7951	0.8978	0.8146	**0.8542**
		DeepSMILES	0.6201	0.7756	0.8477	0.8477	0.8477
	BBBP	SMILES	0.6045	0.7268	0.7387	0.9735	0.8400
		DeepSMILES	0.6098	0.7366	0.7366	1.0000	**0.8483**
	ClinTox	SMILES	0.4795	0.9195	0.9257	0.9928	0.9580
		DeepSMILES	0.4298	0.9262	0.9262	1.0000	**0.9617**
	Tox21^{c_w}	SMILES	0.6814	0.7903	0.9427	0.8270	**0.8811**
		DeepSMILES	0.6831	0.7223	0.9462	0.7469	0.8348
	Tox21	SMILES	0.3563	0.9365	0.9393	0.9969	**0.9672**
		DeepSMILES	0.3148	0.9365	0.9393	0.9969	**0.9672**

Bold values denote the best-achieved performance for clarity and emphasis

c_w class-weighted function, *DeepSMILES* zero-shot learning analysis of BERT

the dataset. The FreeSolv dataset, which focuses on solvation free energy prediction, poses unique challenges due to; Small Dataset Size: FreeSolv contains a relatively small number of samples compared to other molecular datasets, making it harder for our models to generalize effectively.

Inherent Complexity of Solvation Free Energy: Predicting solvation free energy involves intricate intermolecular interactions, which may not be captured adequately by the BERT model's architecture or learned embeddings without domain-specific augmentations.

K-fold cross-validation

Alternative data-splitting strategies could be considered to assess the robustness and generalizability of the model. We conducted experiments using a diverse data-splitting strategy, leveraging K-fold cross-validation on the newly proposed datasets as shown in Table 10. Specifically, the dataset was partitioned into five consecutive folds (K=5) with shuffling to ensure representative subsets across the splits. This methodology helps minimize bias and variance in the performance evaluation by training the model iteratively on four folds while validating on the fifth, rotating through all folds. The experiments employed the our proposed model with only "absolute" and "relative_key_query" PEs, optimized for predicting molecular properties. Herein, significant improvement patterns were observed on both PEs for few of the newly proposed datasets (e.g., COVID and Cocrystals), except for the Malaria dataset, as shown in Table 10.

Molecule splitters (ScaffoldSplitter)

To further evaluate the performance of molecular representations in our study, we employed the ScaffoldSplitter strategy, a specialized tool from DeepChem designed to split datasets based on the molecular scaffolds of small molecules [55]. This approach leverages the Bemis-Murcko scaffold framework to group molecules with similar core structures. While ScaffoldSplitter is optimized for standard SMILES strings, it does not natively support the alternative encoding style introduced by DeepSMILES, which requires preprocessing for compatibility. In this study, we applied ScaffoldSplitter exclusively to SMILES representations, with the results presented in Table 11. Interestingly, compared to other data-splitting strategies utilized in this work, the scaffold-based splitting approach demonstrated lower predictive performance across both classification tasks (i.e., on newly proposed datasets) and regression tasks. These findings suggest that scaffold-based splitting may pose additional challenges for learning models when applied to the datasets considered in this study.

Table 9 Summary on F1 score and RMSE comparisons among position encoding/PEs using BERT

Task	Data	Sequence	Relative_key_query	Sinusoidal	Relative_key	Absolute
Classification (F1 score)	Malaria	SMILES	**0.7439**	0.6811	0.7254	0.6685
		DeepSMILES	0.6532	0.6866	**0.6913**	0.6346
	COVID	SMILES	0.7568	0.7805	**0.8000**	0.7733
		DeepSMILES	0.7568	0.7671	**0.8180**	0.7733
	COVID-19	SMILES	0.7718	0.7950	0.7819	**0.8049**
		DeepSMILES	0.7179	0.7417	**0.7654**	0.7407
	Cocrystals	SMILES	0.5538	**0.6713**	0.6466	0.5538
		DeepSMILES	0.5755	**0.7285**	0.6711	0.5152
	BBBP^{c_w}	SMILES	0.8092	**0.8967**	0.8498	0.8542
		DeepSMILES	0.8456	0.7626	**0.8512**	0.8477
	BBBP	SMILES	0.8555	0.8531	**0.9061**	0.8400
		DeepSMILES	0.8483	0.8473	**0.9196**	0.8483
	ClinTox	SMILES	**0.9617**	0.9577	0.8800	0.9580
		DeepSMILES	**0.9617**	**0.9617**	0.9231	**0.9617**
	Tox21^{c_w}	SMILES	**0.9493**	0.9058	–	0.8811
		DeepSMILES	0.9139	**0.9647**	–	0.8348
	Tox21	SMILES	**0.9688**	0.9680	0.9680	0.9672
		DeepSMILES	**0.9688**	0.9680	0.9680	0.9672
Regression (RMSE)	ESOL	SMILES	0.6185	**0.5883**	0.7878	0.5983
		DeepSMILES	0.6557	0.6256	0.8431	**0.5584**
	FreeSolv	SMILES	**1.8858**	2.0491	2.6242	2.4169
		DeepSMILES	2.1103	2.1572	2.0209	**1.9840**
	Lipophilicity	SMILES	**0.5704**	0.5732	0.5716	0.6025
		DeepSMILES	**0.6333**	0.6707	0.6857	0.6747

Bold values denote the best-achieved performance for clarity and emphasis

c_w class-weighted function, *DeepSMILES* zero-shot learning analysis of BERT, *F1 score* classification tasks, *RMSE* regression tasks

Result on polaris benchmark

To address the limitations associated with the MoleculeNet dataset, which has been criticized for not being representative of real-world datasets [56], we conducted experiments on a selection of fine-tuning benchmarks from improved dataset sources, such as Polaris [57], for comparison as shown in Table 12. The selected datasets were initially split into training and test sets from Polaris. Compared to the MoleculeNet benchmark, notable performance improvements were observed on the BBBP dataset for both PEs (i.e., Absolute and Relative_key_query).

Newly proposed datasets

On the newly proposed datasets, our inherited BERT model also exhibited similar or comparable performance when fine-tuned using SMILES and zero-shot learning, as shown in Tables 4, 6, 7, and 8. During the zero-shot learning analysis (i.e., DeepSMILES representations) most PEs employed in our inherited model showed performance improvements over SMILES representations in at least two of the four tasks, though the magnitude of improvement varied by task. Despite differences in performance across various tasks, the BERT model consistently seemed to handle best the COVID and COVID-19 datasets, indicating its strong capability to handle complex, real-world data related to viral outbreaks.

However, BERT exhibited the lowest F1-score with all PEs except for sinusoidal PE when applied to the cocrystals dataset, suggesting that certain chemical structures might pose a greater challenge for the model. This may also be due to the nature of the cocrystals dataset fed into BERT involved two compounds, namely API and coformers, resulting in a long single sequence compared with the other datasets comprising a single compound. This increased sequence length may posed unique challenges for tokenization and encoding, potentially affecting the model's ability to capture the interactions and relationships between the two components effectively. Therefore, these findings show that, in this field PEs might still struggle with long input sequences, which impact their performance on downstream tasks. This highlights the

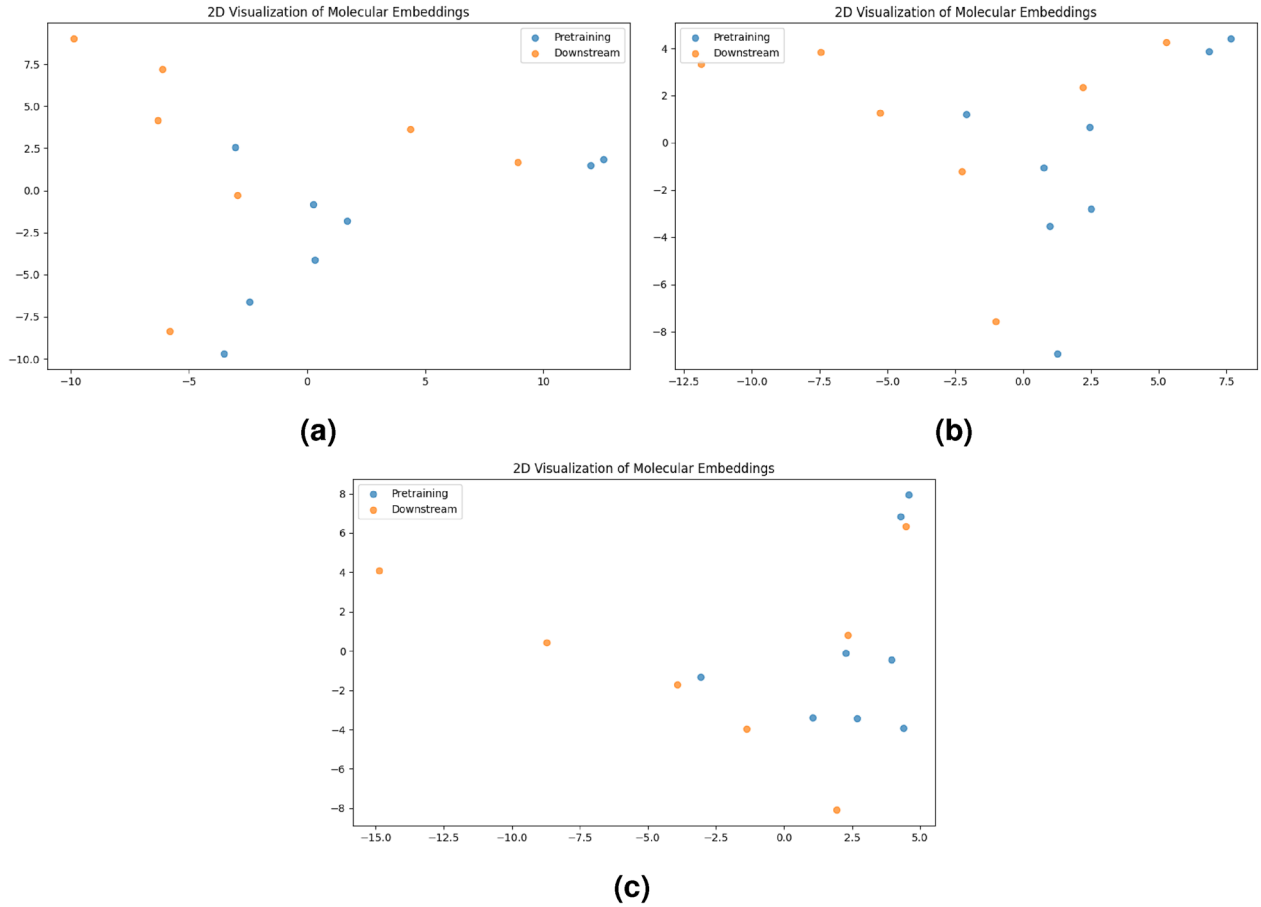


Fig. 3 Analysis of the similarity between the pretraining and downstream (i.e., FreeSolv) SMILES dataset samples (i.e., 8 randomly selected samples) based on the molecular structures. PCA Visualization of molecular embeddings in latent space using BERT model with **a** Absolute, **b** Relative_key_query, **c** Relative_key PEs

Table 10 Averaged performance of fine-tuned BERT using K-fold cross-validation on newly proposed datasets

PE	Data	Sequence	Aver. loss	Aver. accuracy	Aver. precision	Aver. recall	Aver. F1-score
Absolute	Malaria	SMILES	0.5364 ± 0.01	0.7332 ± 0.01	0.7227 ± 0.04	0.6203 ± 0.01	**0.6672 ± 0.02**
		DeepSMILES	0.5662 ± 0.01	0.7130 ± 0.02	0.7189 ± 0.02	0.5510 ± 0.02	0.6234 ± 0.02
	COVID	SMILES	0.5640 ± 0.02	0.7973 ± 0.04	0.8472 ± 0.06	0.7530 ± 0.06	**0.7959 ± 0.05**
		DeepSMILES	0.5316 ± 0.01	0.7716 ± 0.04	0.8010 ± 0.04	0.7560 ± 0.05	0.7770 ± 0.04
	Cocrystals	SMILES	0.5795 ± 0.01	0.6953 ± 0.02	0.6888 ± 0.03	0.6286v0.04	**0.6560 ± 0.02**
		DeepSMILES	0.5723 ± 0.02	0.6807 ± 0.03	0.6692 ± 0.04	0.6188 ± 0.05	0.6418 ± 0.03
RKQ	Malaria	SMILES	0.5025 ± 0.01	0.7616 ± 0.01	0.7542 ± 0.02	0.6647 ± 0.02	**0.7062 ± 0.01**
		DeepSMILES	0.5546 ± 0.01	0.7263 ± 0.01	0.7669 ± 0.04	0.5288 ± 0.04	0.6244 ± 0.02
	COVID	SMILES	0.4777 ± 0.02	0.8432 ± 0.02	0.9079 ± 0.03	0.7818 ± 0.03	**0.8398 ± 0.03**
		DeepSMILES	0.4408 ± 0.04	0.8068 ± 0.03	0.8752 ± 0.01	0.7385 ± 0.06	0.8002 ± 0.04
	Cocrystals	SMILES	0.5247 ± 0.03	0.7197 ± 0.02	0.7127 ± 0.02	0.6633 ± 0.03	0.6867 ± 0.02
		DeepSMILES	0.5154 ± 0.03	0.7416 ± 0.02	0.7323 ± 0.03	0.7041 ± 0.07	**0.7151 ± 0.03**

Bold values denote the best-achieved performance for clarity and emphasis

Aver. Averaged, *RKQ* Relative_key_query, *DeepSMILES* zero-shot learning analysis of BERT

Table 11 Averaged performance of fine-tuned BERT using ScaffoldSplitter on newly proposed and regression datasets

Sequence	PE	Data^{cls}	Aver. loss	Aver. accuracy	Aver. precision	Aver. recall	Aver. F1-score
SMILES	Absolute	Malaria	0.6195	0.6696	0.6622	0.5883	0.5728
		COVID	0.5894	0.7608	0.8013	0.7311	0.7617
		Cocrystals	0.7413	0.4877	0.5179	0.2464	0.2719
	RKQ	Malaria	0.5942	0.6898	0.6902	0.6018	0.5939
		COVID	0.5028	0.8230	0.8981	0.7414	0.8100
		Cocrystals	0.7228	0.6060	0.6990	0.3939	0.4521
Sequence	**PE**	**Data^{reg}**	**Aver. loss**	**Aver. RMSE**			
SMILES	Absolute	ESOL	0.8908	0.9366			
		FreeSolv	8.8316	2.7923			
		Lipophilicity	0.4697	0.6840			
	RKQ	ESOL	1.1851	1.0529			
		FreeSolv	4.9600	2.1698			
		Lipophilicity	0.4666	0.6826			

Aver.: Averaged; *RKQ*: Relative_key_query; *cls*: Classification; *reg*: Regression

Table 12 Performance of fine-tuned BERT on polaris benchmark datasets

Sequence	Data^{cls}	PE	Loss	Accuracy	Precision	Recall	F1-score
SMILES	BBBP	Absolute	0.5189	0.8054	0.8074	0.9970	0.8922
		RKQ	0.4732	0.8079	0.8079	1.0000	0.8937
	Tox21	Absolute	0.3327	0.8952	0.8952	1.0000	0.9447
		RKQ	0.3111	0.8952	0.8952	1.0000	0.9447
Sequence	**Data^{reg}**	**PE**	**Loss**	**RMSE**			
SMILES	Lipophilicity	Absolute	0.4573	0.6763			
		RKQ	0.4297	0.6555			

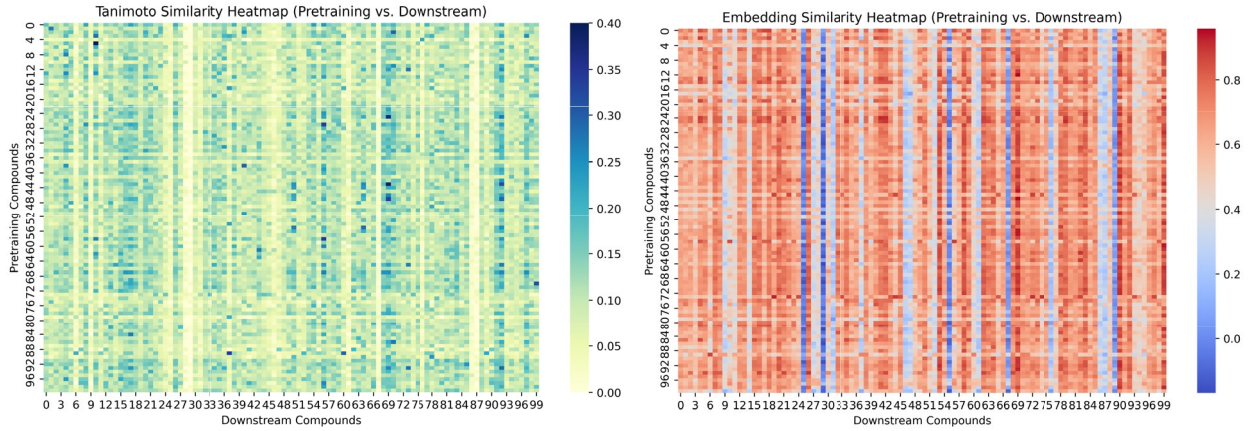
RKQ Relative_key_query, *cls* Classification, *reg* Regression

potential of exploring and selecting appropriate tokenization and PE strategies customized to specific tasks for optimal model performance.

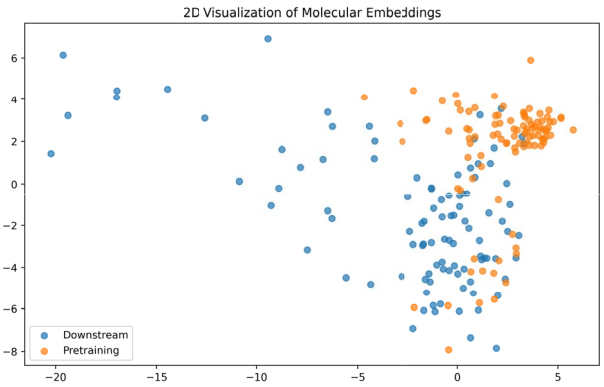
Physiology datasets

On the physiology dataset, BERT with all PEs strongly performed well with and without class weight function, particularly with sinusoidal PE during fine-tuning, as shown in Supplemental Figure S3 and S4. This might be due to the following reasons: Structural Similarity Impact: If the chemical compounds used during pretraining are structurally similar to those in downstream datasets, the pretrained BERT model may exhibit better performance. This is because the model has already learned to represent patterns, relationships, and features present in similar chemical structures during pretraining. Consequently, the embeddings and attention mechanisms of the model can directly transfer this knowledge to downstream tasks, improving predictive accuracy; Semantic Similarity in Text Representations: Physiology-related datasets might encode information (e.g., functional groups, chemical interactions, or pharmacological properties) that align closely with textual patterns seen in pretraining corpora, such as chemical databases or SMILES/DeepSMILES notations. When this overlap occurs, the model benefits from reduced domain discrepancy, as it doesn't need to generalize across entirely unseen contexts.

We perform experimental analyses using three PEs (i.e., "absolute", "relative_key_query", and "relative_key") to derive insights regarding how well the pretraining SMILES data aligns with the downstream SMILES datasets in terms of structural and latent features as shown in Figs. 4, 5, and 6. If the datasets show high similarity, it indicates that the pretrained model was well-suited for downstream tasks. Conversely, low similarity could highlight a potential domain gap requiring additional



(a) Tanimoto Similarity Heatmap **(b)** Embedding Similarity Heatmap

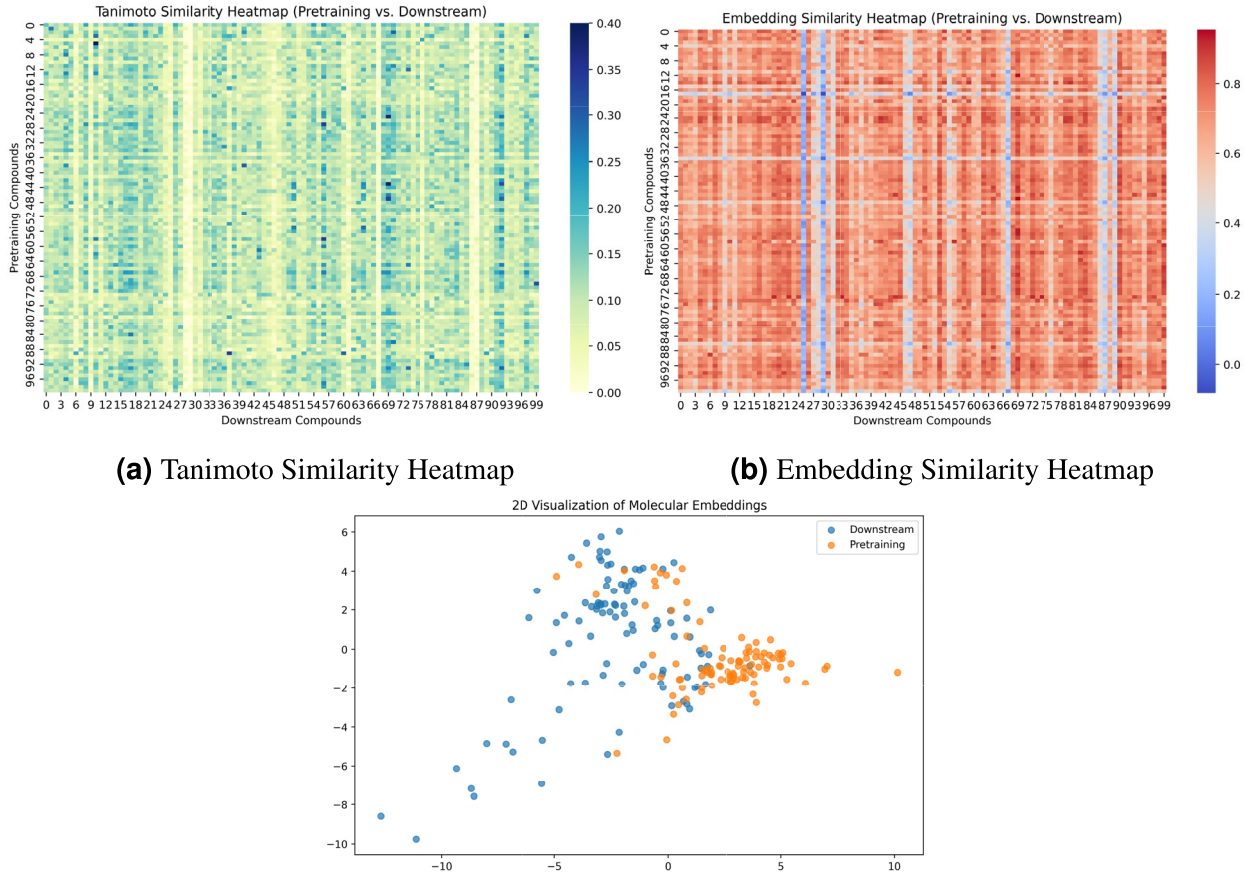


(c) PCA Visualization of Molecular Embeddings

Fig. 4 Analysis of the similarity between the pretraining and downstream (i.e., Clintox) SMILES dataset samples (i.e., 100 randomly selected samples) based on the molecular structures using BERT with "*Absolute*" PE. **a** Tanimoto similarity heatmap, **b** Embedding similarity heatmap, and **c** PCA Visualization of molecular embeddings in latent space

domain-specific fine-tuning. Similarity analysis between the pretraining and downstream SMILES dataset samples was performed based on the molecular structures using techniques like structural similarity measures (e.g., Tanimoto score or Morgan fingerprints) and embedding similarity in latent spaces as follows;

- Figures 4a, 5a, and 6a represent the Tanimoto Similarity Heatmaps for BERT model with "absolute", "relative_key_query", and "relative_key" PEs respectively. These heatmaps show a matrix of similarity scores between pretraining and downstream molecular SMILES samples, based on the Tanimoto coefficient derived from Morgan fingerprints. Each cell in the matrix will have a value between 0 and 1, where 1 indicates identical structural fingerprints and 0 means no similarity. Higher values suggest stronger structural similarity between molecules from the two datasets.
- Figures 4b, 5b, and 6b represent the Embedding Similarity Heatmaps for BERT model with "absolute", "relative_key_query", and "relative_key" PEs respectively. These heatmaps represent the cosine similarity between the embeddings of the pretraining and downstream SMILES samples. Each value represents how similar the latent space representations of the molecules are, with values close to 1 meaning very similar molecular representations (based on BERT's learned embeddings), and values close to 0 suggesting dissimilarity. This provides a



(a) Tanimoto Similarity Heatmap

(b) Embedding Similarity Heatmap

(c) PCA Visualization of Molecular Embeddings

Fig. 5 Analysis of the similarity between the pretraining and downstream (i.e., Clintox) SMILES dataset samples (i.e., 100 randomly selected samples) based on the molecular structures using BERT with "*Relative_key_query*" PE. **a** Tanimoto similarity heatmap, **b** Embedding similarity heatmap, and **c** PCA Visualization of molecular embeddings in latent space

deeper, contextual understanding of how similar the datasets are in terms of molecular features.

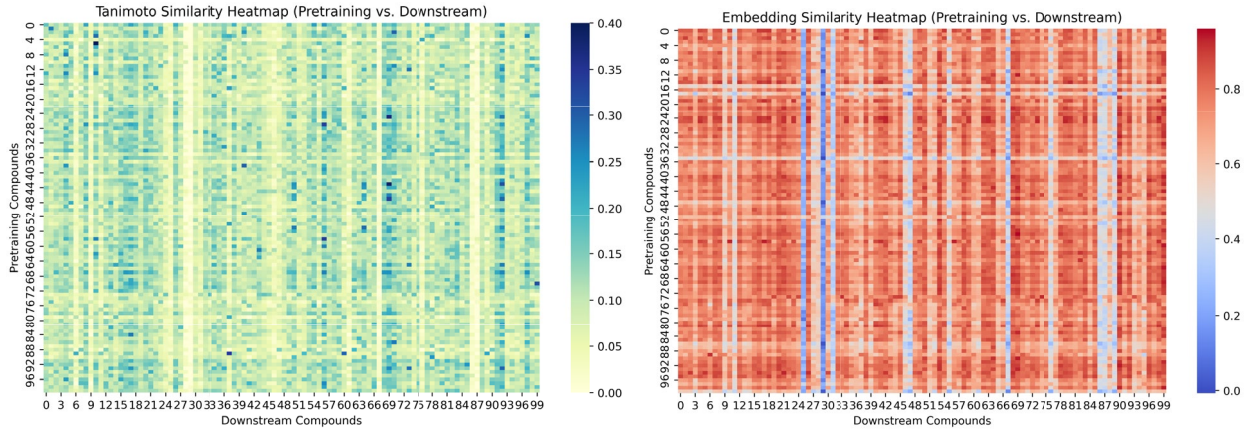
- Figures 4c, 5c, and 6c also represent 2D principal component analysis (PCA) Plot for BERT model with three PEs respectively. In this plot, pretraining compounds and downstream compounds are plotted in a 2D space, and the molecules are reduced to two principal components using PCA to visualize the relationships between them. Points that are closer together indicate higher similarity (structural or embedding-based). Points that are far apart suggest significant differences between samples.

Physical chemistry datasets

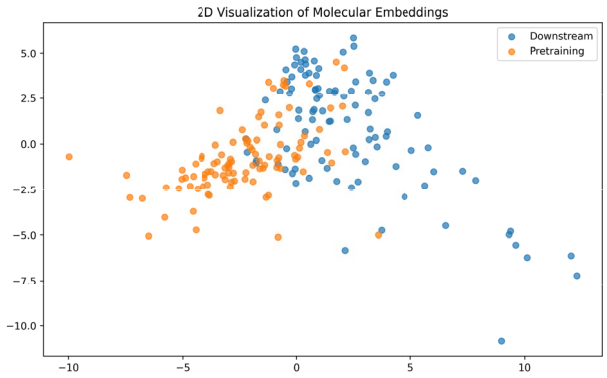
From Tables 4, 6, 7, and 8, we can observe that all PEs in a pretrained BERT model exhibit similar low RMSEs on the ESOL and Lipophilicity datasets but consistently higher RMSEs on the FreeSolv dataset. The following factors could be contributing to this pattern.

Uniform Performance Across PEs: The fact that all PEs result in similar RMSEs suggests that the PE type might not be the primary factor influencing model performance on these datasets. This could mean that the nature of the task is not heavily dependent on the nuances of PEs. Instead, it relies more on the model's ability to understand the overall structure and relationships within the SMILES and DeepSMILES data.

Training Data Coverage: If the pretraining or fine-tuning data has better coverage or representation of the types of molecules found in ESOL and Lipophilicity, the model will naturally perform better on these datasets. Conversely, the model will struggle more if FreeSolv contains data points that are less well-represented or more diverse compared with the training data, resulting in higher errors.



(a) Tanimoto Similarity Heatmap **(b)** Embedding Similarity Heatmap



(c) PCA Visualization of Molecular Embeddings

Fig. 6 Analysis of the similarity between the pretraining and downstream (i.e., Clintox) SMILES dataset samples (i.e., 100 randomly selected samples) based on the molecular structures using BERT with "*Relative_key*" PE. **a** Tanimoto similarity heatmap, **b** Embedding similarity heatmap, and **c** PCA Visualization of molecular embeddings in latent space

Discussion

Initially, we hypothesized that DeepSMILES representations would lead to less robust outcomes as the BERT for each PE was pretrained using only SMILES representations. However, our findings indicated slightly lower differences in downstream performance between the SMILES and DeepSMILES representations for all PEs across various classification and regression tasks. Both the accuracy and F1-score were higher for most tasks when using DeepSMILES compared to SMILES. Furthermore, the test loss was generally lower for tasks using DeepSMILES, indicating improved model performance. Our findings suggest that the proposed BERT model, fine-tuned with various positional encodings and PEs, demonstrates competitive performance across both classification and regression tasks, with certain PEs (such as relative_key_query) showing advantages in specific contexts particularly in zero-shot learning scenarios. This consistent performance demonstrates the BERT model's ability to generalize well across unseen data and highlights the potential of advanced PEs in enhancing the performance of pretrained LMs on chemical and bioactivity prediction tasks.

Limitations

Despite these promising results, this study has several limitations that are important to consider. Firstly, the datasets used especially the fine-tuning datasets from MoleculeNet [7], may not encompass the complete diversity of chemical structures encountered in real-world applications. This limitation could impact the generalizability of the findings, as the model's performance might not be fully representative of the complexities present in real-world chemical data [56].

Another limitation is the evaluation of molecular representations, which was limited to SMILES and DeepSMILES for zero-shot learning. This suggests the need to explore other representations such as SELFIES or InChI

to ensure a more comprehensive assessment of molecular representations and their impact on model performance.

Furthermore, the study used the same tokenization algorithm for SMILES and DeepSMILES, indicating that different tokenization strategies could be examined for potential improvements. Exploring alternative tokenization strategies could provide valuable insights into the impact of tokenization on model performance and generalizability.

Additionally, the computational complexity associated with BERT models necessitates considerations for practical implementation in resource-constrained environments. The study mentions that the complexity of the model during pretraining necessitated the use of a smaller batch size and a reduction in the number of training epochs as observed in Tables 2 and 3. This limitation highlights the need to address the computational overhead of BERT-based models without compromising accuracy, especially in practical applications.

Overall, these limitations underscore the need for further research to address the diversity of datasets, explore alternative molecular representations and tokenization strategies, and optimize the practical implementation of BERT-based models in resource-constrained environments. Addressing these limitations can enhance the comprehensiveness and applicability of the study's findings.

Finally, we compared our work with other ML models reported in previous studies for classification tasks using two newly proposed datasets (i.e., malaria and cocrystals). For the malaria dataset, we implemented a BERT model incorporating the relative key-query PE, showcasing its adaptability for sequence-based data. In contrast, for the cocrystal dataset, we utilized a BERT model with sinusoidal positional encoding, which has been a standard approach in transformer architectures. The comparative results, summarized in Table 13, highlight the performance differences between these models and provide insights into the impact of different positional encodings on dataset-specific tasks.

From the Table, it can be concluded that the BERT-based models utilizing SMILES and DeepSMILES sequences demonstrate competitive performance compared to traditional ML models. While the accuracy of BERT with SMILES and DeepSMILES sequences in both datasets is lower than graph neural networks (GNNs) models, the use of sequence-based inputs reflects the flexibility of BERT in handling text-based chemical representations. This highlights the potential of transformer-based architectures like BERT in cheminformatics tasks, particularly for applications where pretraining on chemical sequences might provide broader generalization capabilities compared to traditional feature engineering approaches.

Conclusion and future work

In classification tasks, the BERT model demonstrates lower performance on newly proposed datasets regardless of the PEs, such as the cocrystals dataset [40]. However, the model utilizing DeepSMILES occasionally showed performance improvement compared to

Table 13 Performance Comparison of ML models from previous studies for Classification Using different feature representations versus our study (i.e., BERT with different positional encoding strategies)

Dataset	Study	Features	Model	Accurary
Malaria	Mswahili et al. [58]	PaDEL	SVM/LR	0.7850/0.7795
			RF/ANN	0.8294/0.8223
	Egieyeh et al.[59]	RDKit	SVM	0.8593
	Danishuddin et al. [60]	PaDEL	SVM & XGBoost	0.8500
	Mswahili et al. [39]	Mordred/BERT	RGCN	**0.9958/0.9958**
	Our work	SMILES sequences	BERT (RKQ)	0.8017
		DeepSMILES sequences	BERT (RKQ)	0.7495
Cocrystals	Wicker et al. [61]	RDKit	SVM	0.6400
	Mswahili et al. [40]	Mordred	ANN/RF	0.8330/0.8290
			XGB/SVM	0.8320/0.7460
	Devogealer et al. [62]	ECFP	ANN	0.8000
	Mswahili et al. [43]	Mordred	RGCN/GCN	**0.9595/0.9136**
	Our work	SMILES sequences	BERT (Sinusoidal PE)	0.7134
		DeepSMILES sequences	BERT (Sinusoidal PE)	0.7500

Bold values denote the best-achieved performance for clarity and emphasis

DeepSMILES zero-shot learning analysis of BERT, *ECFP* Extended-connectivity fingerprints, *SVM* Support Vector Machine, *LR* Logistic Regression, *RF* Random Forest, *ANN* Artificial Neural Network, *XGB* Extreme Gradient Boosting, *RGCN* Relational Graph Convolution Network, *GCN* Graph Convolution Network

SMILES in terms of accuracy and F1-score. This suggests that the DeepSMILES representation captures more relevant information for these tasks, leading to improved performance. The lower test loss observed when using DeepSMILES across most tasks further corroborates its effectiveness. Notably, the relative_key_query PE method demonstrates slightly better performance compared to the other PEs, indicating its remarkable capability to capture dependencies within SMILES and DeepSMILES input sequences.

For regression tasks, the performance tends to vary. Although DeepSMILES generally outperforms SMILES, particularly for the FreeSolv dataset, the performance difference is less pronounced on the ESOL and Lipophilicity dataset. The significant test loss and RMSE increase on the FreeSolv dataset using SMILES suggest that the model struggles with this representation in specific contexts, emphasizing the importance of choosing an appropriate molecular representation for different tasks.

Future research should address these limitations by expanding the range of molecular representations and datasets to confirm the generalizability of the findings. As observed during training, Table 3 suggests that further investigations into reducing the computational overhead of BERT-based models without compromising accuracy could enhance their practical utility. Exploring hybrid approaches that combine BERT with other machine learning techniques or domain-specific knowledge may also yield notable improvements. Long-term observational research evaluating the real-world impact of these models on drug discovery and material science applications would provide valuable insights into their practical benefits.

Additionally, several tokenizers are available in cheminformatics, such as SMILESTokenizer [23], Atom-in-SMILES [63], SMILES pair encoding imitating BPE [64], Atom-level tokenization [64], k-mer (also known as n-grams) tokenization [64], and BPE [3, 23]. Of these, we adopted the SmilesTokenizer module from DeepChem [45, 46]. While other tokenizers were not considered in this study, they will be explored for comparison purposes in future research, provided that tokenization considerably impacts quality of predictions in text generation frameworks [63], such as SMILES-based predictions [65]. In this case, we aim to not only explore the tokenization algorithms but also to investigate the effects and contributions of individual tokens, including branches, bonds, and the first, middle, and last elements from both SMILES and DeepSMILES representations.

Future work

In the near future, we aim to focus our research efforts on addressing the following critical issues. This will involve an in-depth exploration of the underlying challenges, a thorough evaluation of existing methodologies and datasets, and the development of innovative solutions.

- We aim to extend our research by exploring alternative PE techniques, with a particular focus on Rotary Position Embedding (RoPE) to effectively leverage the positional information [29]. RoPE has shown potential in improving the representation of sequential data, which may lead to more efficient encoding of chemical structures.
- Additionally, we plan to extend this study to gain a deeper understanding of the tokenization process, specifically investigating the contribution of individual tokens in both SMILES and DeepSMILES representations. With an improved pretraining datasets, this analysis could provide valuable insights into how tokenization influences model performance and chemical related task predictions.
- We intend to incorporate zero-shot learning mechanism to evaluate the model's ability to generalize across unseen data, assessing its robustness and adaptability to new tasks or domains.
- Furthermore to address the flaws associated with MoleculeNet dataset like to be considered not representative of real-world data set [56], we anticipate in utilizing other fine-tuning benchmarks from other improved datasets sources such as Polaris [57], MetaMolNet [66].

We believe these extensions will not only broaden the scope of our research but also to contribute meaningful advancements to the field and bridge existing gaps in knowledge or application of NLP LMs in cheminformatics.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00959-9.

Supplementary Material 1.

Acknowledgements

Data analysis and Artificial Intelligence (DAI) Lab at Chungbuk National University, South Korea & Health Informatics Lab (HIL) at Nanyang Technological University, Singapore.

Author contributions

M.E.M., and Y.-S.J. conceptualization; M.E.M. training data curation; M.E.M., and J.H. fine-tuning data curation; M.E.M., and J.H. conceived the experiments; M.E.M., and J.H. conducted the experiments; M.E.M., and J.H. analysed the results; M.E.M. wrote the original manuscript; Y.-S.J., J.C.R., and K.J. reviewed the manuscript; Y.-S.J. supervision. All authors reviewed the manuscript.

Funding

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A3053015). This work was supported by the National

Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00217022).

Availability of data and materials
The dataset and source codes utilized in this study are accessible online at: https://github.com/medard2505/Positional-Embeddings-and-Zero-shot-Learning-using-BERT-for-Molecular-Property-Prediction

Declarations

Ethics approval and consent to participate
This study does NOT involve any human or animal subjects.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Computer Engineering, Chungbuk National University, Cheongju 28644, South Korea. [2]School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore.

Received: 19 September 2024 Accepted: 18 January 2025
Published online: 05 February 2025

References

1. Li J, Jiang X (2021) Mol-bert: an effective molecular representation with bert for molecular property prediction. Wirel Commun Mobile Comput 2021:1–7
2. Liu Y et al (2023) Molrope-bert: an enhanced molecular representation with rotary position embedding for molecular property prediction. J Mol Graph Model 118:108344
3. Lee I, Nam H (2022) Infusing linguistic knowledge of smiles into chemical language models. arXiv preprint arXiv:2205.00084
4. Vamathevan J et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18:463–477
5. Ekins S et al (2019) Exploiting machine learning for end-to-end drug discovery and development. Nat Mater 18:435–441
6. Walters WP, Barzilay R (2020) Applications of deep learning in molecule generation and molecular property prediction. Acc Chem Res 54:263–270
7. Wu Z et al (2018) Moleculenet: a benchmark for molecular machine learning. Chem Sci 9:513–530
8. Thakkar A, Kogej T, Reymond J-L, Engkvist O, Bjerrum EJ (2020) Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. Chem Sci 11:154–168
9. Abdel-Aty H, Gould IR (2022) Large-scale distributed training of transformers for chemical fingerprinting. J Chem Inf Model 62:4852–4862
10. Winter R, Montanari F, Noé F, Clevert D-A (2019) Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci 10:1692–1701
11. Sadybekov AV, Katritch V (2023) Computational approaches streamlining drug discovery. Nature 616:673–685
12. Arús-Pous J et al (2019) Exploring the GDB-13 chemical space using deep generative models. J Cheminform 11:1–14
13. Fabian B et al (2020) Molecular representation learning with language models and domain-relevant auxiliary tasks. arXiv preprint arXiv:2011.13230
14. Chen D, Zheng J, Wei G-W, Pan F (2021) Extracting predictive representations from hundreds of millions of molecules. J Phys Chem Lett 12:10793–10801
15. Balaji S, Magar R, Jadhav Y et al (2023) Gpt-molberta: Gpt molecular features language model for molecular property prediction. arXiv preprint arXiv:2310.03030
16. Grisoni F (2023) Chemical language models for de novo drug design: challenges and opportunities. Curr Opin Struct Biol 79:102527
17. Dai AM, Le QV (2015) Semi-supervised sequence learning. Advances in neural information processing systems; 2015. p. 28
18. Yu L, Su Y, Liu Y, Zeng X (2021) Review of unsupervised pretraining strategies for molecules representation. Brief Funct Genomics 20:323–332
19. Liu Z et al (2021) Ai-based language models powering drug discovery and development. Drug Discov Today 26:2593–2607
20. Xu Z, Wang S, Zhu F, Huang J (2017) 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. pp 285–294
21. Zhang X et al (2018) 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. pp 404–413
22. Wang S, Guo Y, Wang Y, Sun H, Huang J (2019) 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. pp 429–436
23. Chithrananda S, Grand G, Ramsundar B (2020) Chemberta: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885
24. Ahmad W, Simon E, Chithrananda S, Grand G, Ramsundar B (2022) Chemberta-2: towards chemical foundation models. arXiv preprint arXiv:2209.01712
25. Zhang C et al (2024) Transfer learning across different chemical domains: virtual screening of organic materials with deep learning models pretrained on small molecule and chemical reaction data. J Cheminform 16:89
26. Li B, Lin M, Chen T, Wang L (2023) Fg-bert: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. Brief Bioinform 24:bbad398
27. Zhang X-C et al (2021) Mg-bert: leveraging unsupervised atomic representation learning for molecular property prediction. Brief Bioinform 22:bbab152
28. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
29. Su J et al (2024) Roformer: enhanced transformer with rotary position embedding. Neurocomputing 568:127063
30. O'Boyle N, Dalke A (2018) Deepsmiles: an adaptation of smiles for use in machine-learning of chemical structures. ChemRxiv. https://doi.org/10.26434/chemrxiv.7097960.v1
31. Sultan A, Sieg J, Mathea M, Volkamer A (2024) Transformers for molecular property prediction: lessons learned from the past five years. J Chem Inf Model 64:6259–6280
32. Labrak Y, Rouvier M, Dufour R (2023) A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. arXiv preprint arXiv:2307.12114
33. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) Zinc: a free tool to discover chemistry for biology. J Chem Inf Model 52:1757–1768
34. Irwin JJ et al (2020) Zinc20-a free ultralarge-scale chemical database for ligand discovery. J Chem Inf Model 60:6065–6073
35. Kim S et al (2019) Pubchem 2019 update: improved access to chemical data. Nucleic Acids Res 47:D1102–D1109
36. Kim S et al (2023) Pubchem 2023 update. Nucleic Acids Res 51:D1373–D1380
37. Gaulton A et al (2017) The chembl database in 2017. Nucleic Acids Res 45:D945–D954
38. Gaulton A et al (2012) Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
39. Mswahili ME, Ndomba GE, Jo K, Jeong Y-S (2024) Graph neural network and bert model for antimalarial drug predictions using plasmodium potential targets. Appl Sci 14:1472
40. Mswahili ME et al (2021) Cocrystal prediction using machine learning models and descriptors. Appl Sci 11:1323
41. Mswahili ME, Hwang J, Jeong YS, Kim Y (2022) Graph neural network models for chemical compound activeness prediction for covid-19 drugs discovery using lipinski's descriptors. In: 2022 5th international conference on artificial intelligence for industries (AI4I). IEEE. pp 20–21
42. Harigua-Souiai E et al (2021) Deep learning algorithms achieved satisfactory predictions when trained on a novel collection of anticoronavirus molecules. Front Genet 12:744170

43. Mswahili ME, Jo K, Jeong Y-S, Lee S (2024) Graph neural networks with multi-features for predicting cocrystals using APIS and coformers interactions. Curr Med Chem 31:5953–5968
44. Zhu M et al (2018) Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access 6:4641–4652
45. DeepChem tokenizers. https://deepchem.readthedocs.io/en/2.4.0/api_reference/tokenizers.html. Accessed 25 Sept 2023
46. Ramsundar B et al (2016) Democratizing deep-learning for drug discovery, quantum chemistry, materials science and biology. GitHub repository
47. Wolf T et al (2020) Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. pp 38–45
48. Schwaller P et al (2021) Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell 3:144–152
49. Schwaller P et al (2019) Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Central Sci 5:1572–1583
50. Huang Z, Liang D, Xu P, Xiang B (2020) Improve transformer models with better relative position embeddings. arXiv preprint arXiv:2009.13658
51. Qu A, Niu J, Mo S (2021) Explore better relative position embeddings from encoding perspective for transformer models. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp 2989–2997
52. Vaswani A et al (2017) Attention is all you need. Advances in neural information processing systems. p 30
53. Shaw P, Uszkoreit J, Vaswani A (2018) Self-attention with relative position representations. arXiv preprint. arXiv:1803.02155
54. Lv Q, Zhou J, Yang Z, He H, Chen CY-C (2023) 3d graph neural network with few-shot learning for predicting drug-drug interactions in scaffold-based cold start scenario. Neural Netw 165:94–105
55. Molecule Splitters scaffoldsplitter. https://deepchem.readthedocs.io/en/latest/api_reference/splitters.html#scaffoldsplitter. Accessed 26 Dec 2024
56. We need better benchmarks for machine learning in drug discovery. https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html. Accessed 27 Nov 2024
57. The benchmarking platform for drug discovery. https://polarishub.io/. Accessed 27 Nov 2024
58. Mswahili ME, Martin GL, Woo J, Choi GJ, Jeong Y-S (2021) Antimalarial drug predictions using molecular descriptors and machine learning against *Plasmodium falciparum*. Biomolecules 11:1750
59. Egieyeh S, Syce J, Malan SF, Christoffels A (2018) Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach. PLoS ONE 13:e0204644
60. Danishuddin, Madhukar G, Malik M, Subbarao N (2019) Development and rigorous validation of antimalarial predictive models using machine learning approaches. SAR QSAR Environ Res 30:543–560
61. Wicker JG et al (2017) Will they co-crystallize? CrystEngComm 19:5336–5340
62. Devogelaer J-J, Meekes H, Tinnemans P, Vlieg E, De Gelder R (2020) Co-crystal prediction by artificial neural networks. Angew Chem Int Ed 59:21711–21718
63. Ucak UV, Ashyrmamatov I, Lee J (2023) Improving the quality of chemical language model outcomes with atom-in-smiles tokenization. J Cheminform 15:55
64. Li X, Fourches D (2021) Smiles pair encoding: a data-driven substructure tokenization algorithm for deep learning. J Chem Inf Model 61:1560–1569
65. Domingo M, Garcıa-Martınez M, Helle A, Casacuberta F, Herranz M (2018) How much does tokenization affect neural machine translation? arXiv preprint arXiv:1812.08621
66. Lv Q, Chen G, Yang Z, Zhong W, Chen CY-C (2024) Meta-molnet: a cross-domain benchmark for few examples drug discovery. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2024.3359657

Publisher's Note