Smile-to-Bert: A BERT architecture trained for physicochemical properties prediction and SMILES embeddings generation

Maria Barranco-Altirriba,[*,†,‡,¶] Vivian Würf,[§] Enrico Manzini,[†,‡,‡] Josch K. Pauling,[§,‖] and Alexandre Perera-Lluna[†,‡,‡]

†*B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Av. Diagonal, 647, 08028 Barcelona, Spain*

‡*Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), 28029 Madrid, Spain*

¶*Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain*

§*LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany*

‖*CioBio, Institute for Clinical Chemistry and Laboratory Medicine, University Hospital and Faculty of Medicine Carl Gustav Carus, Dresden University of Technology, 01307 Dresden, Germany*

E-mail: maria.barranco@upc.edu

Abstract

Binding affinity and molecular property prediction are crucial for drug discovery. Over the years, deep learning models have been widely used for these tasks; however, large datasets are often needed to achieve strong performances. Pre-training models on vast unlabelled data has emerged as a method to extract contextualised embeddings

that boost performance on smaller datasets. SMILES (Simplified Molecular Input Line Entry System) encode molecular structures as strings, making them suitable for natural language processing (NLP). Transformers, known for capturing long-range dependencies, are well-suited for processing SMILES. One such transformer-based architecture is BERT (Bidirectional Encoder Representations from Transformers), which only uses the encoder part of the Transformer and performs classification and regression tasks. Pre-trained transformer-based architectures using SMILES have significantly improved predictions on smaller datasets. Public data repositories such as PubChem, which provide SMILES and physicochemical properties among other data, are essential for pre-training these models. SMILES embeddings that combine chemical structure and physicochemical properties information could further enhance performance on tasks such as binding affinity prediction. For this reason, we introduce Smile-to-Bert, a pre-trained BERT architecture that predicts seven physicochemical properties from SMILES using PubChem data and two different SMILES tokenizers. Moreover, this model generates embeddings that integrate information about molecular structure and physicochemical properties. Regarding the prediction of physicochemical properties, the mean absolute errors obtained are: H-bond acceptors (0.0502), H-bond donors (0.0048), rotatable bonds (0.0949), exact mass (0.5678), TPSA (0.4961), heavy atom count (0.0345), and log-P (0.2219). Additionally, the usefulness of the generated embeddings is evaluated using two binding affinity datasets and their performance is compared to embeddings created by a state-of-the-art Transformer. We show that the SmilesPE tokenizer works better than the atom-level one and that integrating the embeddings generated by Smile-to-Bert to the state-of-the-art Transformer embeddings improves the prediction of binding affinity in one of the datasets. A dashboard for the prediction of physicochemical properties is available at `http://147.83.252.32:8050/`, and the code is accessible at `https://github.com/m-baralt/smile-to-bert`.

Introduction

Accurate prediction of binding affinity[1] and molecular properties[2] plays a crucial role in computer-aided drug discovery. In recent years, there has been a significant rise in the use of deep learning models for drug discovery, driven by the increasing availability of compound-related data.[3] Public databases such as PubChem,[4] which contain vast amounts of chemical information, exemplify the increasing availability of data in the field. However, acquiring large labelled datasets for binding affinity or molecular property prediction is both challenging and costly, which hinders the training of deep learning models.[1,2] Transfer learning has emerged as a solution to overcome data limitations in deep learning.[1,5] It involves pre-training a model on one data domain and then adapting it to a different domain of interest, which allows the extraction of meaningful features from vast datasets, even when relevant labels are limited.[1]

SMILES (simplified molecular input line entry system) are text-based representations that encode a molecular graph in a simple, human-readable sequence of characters.[6,7] Due to their linear nature, SMILES are widely used in natural language processing (NLP) algorithms.[8] The order of SMILES characters does not always represent the physical proximity of atoms in a molecule, which can complicate the understanding of long-range dependencies. This poses a challenge for recurrent neural networks (RNN) and showcases the usefulness of transformer-based architectures.[1]

The Transformer, first introduced by Vaswani et al.,[9] uses a self-attention mechanism that processes all inputs simultaneously, thus learning long-range dependencies as easily as short-range ones. Using self-attention to process SMILES strings allows for the learning of atomic structure features without relying on assumptions about the sequential nature of the data. Transfer learning has been widely used in Transformer-based architectures. Some decoder-only architectures use unidirectional language models during pre-training, which may limit their ability to learn effectively for certain tasks, such as sentence-level understanding, where context from both directions is crucial. To address these limitations, BERT

(Bidirectional Encoder Representations from Transformers), which only uses the encoder part of the Transformer, was introduced. In the original paper, BERT enables bidirectional context understanding using a masked language modelling as a pre-training objective.[10]

BERT architectures have been pre-trained with the goal of generating contextualised embeddings that can boost the performance of models trained on smaller datasets.[5] More specifically, transformer-based architectures pre-trained through various tasks have been applied to a wide range of applications for processing SMILES. For instance, a Transformer-based architecture using SMILES as input was pre-trained with a masked language modelling objective and subsequently fine-tuned for various classification and regression tasks related to molecular properties. The fine-tuning involved small labelled datasets, where molecular properties such as human clearance or water solubility were predicted.[2] Similarly, to address the limited amount of data in binding affinity assays, a Transformer architecture was pre-trained on the task of translating SMILES strings to IUPAC names. By leveraging the SMILES embeddings obtained by the pre-trained model, binding affinity prediction showed improvement across three datasets compared to using untrained embedding vectors.[1]

Understanding molecules' behaviour in humans or the environment largely depends on its key physicochemical properties. These properties can currently be measured through experimental methods or predicted using quantitative structure-property relationship (QSPR) models. However, experimental methods are often costly, time-consuming, and challenging. Consequently, in silico QSPR modelling provides a valuable alternative by enabling the rapid estimation of physicochemical properties through the identification of relationships between molecular structures and their associated properties.[11]

Given this context, we hypothesise that incorporating physicochemical properties information into embedding vectors might enhance their richness and utility for various SMILES-related tasks, including binding affinity prediction. Therefore, we present Smile-to-BERT, a BERT architecture pre-trained on a large subset of PubChem molecules to predict seven physicochemical properties from SMILES. In this work, we (1) provide a dashboard for pre-

dicting seven physicochemical properties from SMILES, (2) interpret the pre-trained model and its predictions using integrated gradients, (3) evaluate the embeddings generated by our model for binding affinity prediction in two different datasets, (4) compare the binding affinity prediction performance with another pre-trained model, and (5) assess two different SMILES tokenizers.

Materials and Methods

Training and validation data and model architecture

The sdf PubChem files available in `https://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/SDF/` were downloaded using the RCurl R package v1.98.1.14[12] and processed using the ChemmineR package v3.46.0.[13] Compounds with missing values in at least one of the physicochemical properties were discarded and a filter based on the length of the SMILES sequences was applied. Specific details of the input, model architecture and output are described in the following subsections.

Input - SMILES

The first step in NLP algorithms is tokenizing the sequence. Various approaches exist for tokenizing SMILES, but character-level tokenization can lead to issues, such as splitting atoms represented by multiple characters into separate tokens. Atom-level tokenization is a more common approach that ensures each atom is extracted as a distinct token, regardless of the number of characters used to represent that atom. A more sophisticated method is SMILES pair encoding (SmilesPE), a substructural tokenization algorithm inspired by the byte pair encoding algorithm. SmilesPE detects and preserves common SMILES substrings as unique tokens, generating tokens with richer information that more accurately reflect molecular functionalities compared to atom-level tokenization. Additionally, input sequences become shorter compared to those produced by atom-level tokenization.[7] Therefore, we employed an

atom-level tokenizer for SMILES using the `BasicSmilesTokenizer` from `deepchem` python module v2.5.0;[14] and the SmilesPE tokenizer using `SPE_Tokenizer` from `SmilesPE` module.[7]

SMILES sequences longer than 384 and 113 using the atom-level and the SmilesPE tokenizers respectively were removed. The mentioned lengths were selected based on the length of SMILES from LipidMaps,[15,16] as lipids are among the largest biological small molecules, thus ensuring the model's applicability in bioinformatics, particularly in metabolomics and lipidomics. To avoid excessively long sequences, the eight longest lipids (prostaglandins) were discarded due to their significant length difference compared to the rest of lipids' SMILES. We prioritised enforcing a consistent rule for selecting the sequence length threshold over creating a single dataset. As a result, the differing sequence lengths generated by the two tokenizers led to the creation of two datasets with slightly different sample sizes.

Model Architecture

In a BERT model, the encoder section of the Transformer is employed to generate contextualised representations for each input token, which can then be used for downstream tasks, such as predicting a vector of physicochemical properties. Initially, each token in the input sequence is mapped to a dense embedding vector, capturing its semantic meaning. A positional encoding is added to these embeddings to incorporate the sequential information of the tokens. These token embeddings are then passed through a series of stacked encoder layers. Each encoder layer applies multi-head self-attention, which allows capturing complex dependencies between tokens. After this, feed-forward neural networks and normalisation operations are applied. The final contextualised vector serves as input for the specific downstream task.

Our BERT architecture was composed of 4 encoder layers, 6 heads and an embedding size of 90 was used. We used a reduced version of the original BERT to speed up training. A scheme of the architecture can be seen in Figure S1. The embedding size was selected by comparing the performance of 4 different settings with 10M samples (Figure S2). Specific

sequence lengths were enforced by padding the token sequences. A specific amount of the first tokens (95 percentile of the PubChem SMILES lengths) of the final contextualised matrix were concatenated and used to predict seven physicochemical properties using one single linear layer. One architecture for each tokenizer was built, one that handles sequences of 384 tokens and concatenates the first 83 (95 percentile of the PubChem SMILES lengths when using the atom-level tokenizer), and one architecture that accepts sequences of 113 tokens and concatenates the first 30 (95 percentile of the PubChem SMILES lengths when using the SmilesPE tokenizer) to predict physicochemical properties.

Output - Physicochemical properties

Several physicochemical properties were available: complexity, H-bond acceptor count, H-bond donor count, rotatable bond count, exact mass, molecular weight, topological polar surface area (TPSA), total charge, heavy atom count, defined atom stereocentre count, undefined atom stereocentre count, defined bond stereocentre count, undefined bond stereocentre count, isotopic atom count, component count and log-P. From these properties, the molecular weight was discarded because it is redundant with the exact mass, and the total charge, the defined atom stereocentre count, undefined atom stereocentre count, defined bond stereocentre count, undefined bond stereocentre count, the isotopic atom count and the component count were discarded due to the small variability of these properties (Table S1).

Outliers were removed by retaining only compounds that met the following criteria for all physicochemical properties: $Y_i \leq Q3_i + 5IQR_i$ and $Y_i \geq Q1_i - 5IQR_i$, where Y_i is physicochemical property i, $Q1_i$ and $Q3_i$ are the first and third quartiles for Y_i and $IQR_i = Q3_i - Q1_i$.

After this last filter, 99,886,786 compounds remained in the atom-level tokenizer dataset, and 99,885,676 in the SmilesPE tokenizer dataset, from an original amount of 103,281,525 compounds.

Before training, properties values were normalised by subtracting their corresponding median value and dividing by their IQR. This scaling was selected because of its robustness regardless the presence of outliers. Once the properties were scaled, they were multiplied by a constant factor of 100. This last step was performed to avoid the vanishing gradients problem (see Supporting Information for details).

Model training

The models were trained using PyTorch v2.3.0+cu121,[17] and experiments were tracked with wandb v0.16.3.[18] L1Loss function was used to compute the loss and Adam optimiser[19] was applied with an initial learning rate of 1e-4, betas of (0.9, 0.999) and a weight decay of 0.01. The optimiser scheduler used by Vaswani et al[9] was employed. Data was split randomly for training (95%) and validation (5%) and a batch size of 64 was used during training. The models were trained during 20 epochs and the models which obtained the minimum loss in validation were selected. Accelerate v0.27.2[20] was used for distributed training in a total number of 2 virtual machines with 2 NVIDIA A40 GPUs each.

Model performance and interpretation

The validation losses for each property were summarised. Moreover, the relationship between the loss and the property value in the continuous properties was visualised.

Principal Component Analysis (PCA) of the embeddings of 30,400 randomly selected SMILES was conducted using sklearn v1.4.0.[21] The first two principal components were used to visualise the selected embeddings coloured by six physicochemical properties.

Captum module v0.7.0[22] was used to better understand the embedding layer of our models using the integrated gradients algorithm. The importance scores assigned by the embedding layer to each input token when predicting physicochemical properties were computed and visualised for the SMILES of two Pubchem compounds.

Binding affinity prediction

Two datasets were used to assess the utility of the embeddings generated by our BERT models for tasks involving SMILES and to compare their performances with embeddings obtained from a transformer trained to translate SMILES to IUPAC names.[1] We trained and evaluated a binary classifier to predict binding affinity from the embeddings using both datasets. To ensure comparability, we employed the same model architecture as described in the mentioned Transformer paper.[1] The model includes four linear layers with the following dimensions: *embedding vectors size* $\times$ 500, 500 $\times$ 500, 500 $\times$ 100, and 100 $\times$ 2. A ReLU activation function is applied after each layer except the output layer, and a dropout of 0.2 is applied after the second layer.

The datasets were downloaded from `https://github.com/mpcrlab/MolecularTransformerEmbedding` and the non-binding molecules were undersampled to obtain balanced datasets using a function provided in the same Github repository.

The dataset AID 652067[23] contains molecules labelled as binding to the DAF-12 nuclear receptor in H. contortus and molecules that are non-binding. After undersampling, 10700 molecules remained.

The dataset AID 1053197[24] contains molecules labelled as binding when they caused inhibition of sialic acid acetyl-esterase (SIAE) and molecules labelled as non-binding when inhibition was below the threshold specified by the authors. After balancing the dataset, 5110 compounds were kept.

Binding affinity was predicted from trained and untrained embeddings generated by our two BERT architectures and the pre-trained Transformer. The same validation scheme used by Morris et al.[1] was employed. A 10-fold cross validation strategy was used and the mean Receiver Operating Characteristic (ROC) and mean area under the curve (AUC) were computed using sklearn v1.4.0.[21]

The embeddings built using our BERT architectures had considerably smaller sizes than the transformer embeddings, 7,470 and 2,700 for the atom-level and the SmilesPE tokenizer,

respectively, and 131,072 for the transformer. In order to understand the importance of the embedding vectors size and the task used for pre-training, two variations of the models were evaluated. First, a large variation of our BERT architecture using the SmilesPE tokenizer was trained during 10 epochs, with a d_{model} equal to 510 and concatenating all tokens for prediction, thus significantly increasing the size of the embedding vectors to 57,630. Second, the Transformer was fine-tuned during one epoch to predict physicochemical properties using our training data and then, the embeddings built with this model were evaluated for binding affinity prediction.

Finally, to determine if integrating physicochemical properties knowledge enhanced the embedding vectors for the binding affinity prediction task, we combined the embeddings from the pre-trained transformer with those from the large BERT architecture and evaluated them on both binding affinity datasets.

Results

Data description

Depending on the tokenization used, we obtain slightly different descriptive statistics of the data used for training and validation. Table 1 shows the descriptive statistics of the data for both tokenizers.

Table 1: Descriptive statistics of PubChem data for both tokenizers. A total number of 99,886,786 and 99,885,676 compounds were used for training and validation in the atom-level and SmilesPE cases, respectively.

	Tokenizer	Mean	SD	Min	25%	50%	75%	Max
H-bond Acceptor	Atom-level	4.8459	2.3464	0.0000	3.0000	5.0000	6.0000	20.0000
	SmilesPE	4.8459	2.3464	0.0000	3.0000	5.0000	6.0000	20.0000
H-bond Donor	Atom-level	1.2971	1.0564	0.0000	1.0000	1.0000	2.0000	6.0000
	SmilesPE	1.2972	1.0564	0.0000	1.0000	1.0000	2.0000	6.0000
Rotatable Bond	Atom-level	5.5698	3.1351	0.0000	3.0000	5.0000	7.0000	21.0000
	SmilesPE	5.5698	3.1351	0.0000	3.0000	5.0000	7.0000	21.0000
Exact Mass	Atom-level	363.4524	125.5739	1.0078	275.1223	343.9796	429.1627	1255.2079
	SmilesPE	363.4464	125.5591	1.0078	275.1205	343.9789	429.1623	1255.1671
TPSA	Atom-level	76.3140	38.1295	0.0000	50.1000	72.7000	97.6000	346.0000
	SmilesPE	76.3142	38.1294	0.0000	50.1000	72.7000	97.6000	346.0000
Heavy Atom Count	Atom-level	25.2392	9.1625	0.0000	19.0000	24.0000	30.0000	90.0000
	SmilesPE	25.2385	9.1610	0.0000	19.0000	24.0000	30.0000	90.0000
Log-P	Atom-level	3.1884	2.2944	-10.7000	1.8000	3.0000	4.3000	16.8000
	SmilesPE	3.1883	2.2941	-10.7000	1.8000	3.0000	4.3000	16.8000

Model performance and interpretation

The general loss and the loss for each property is summarised in Table 2 for both variations of Smile-to-Bert. The learning curves are shown in Figure S3.

The number of samples used for validation is 4,994,340 and 4,994,284 when using the atom-level and the SmilesPE tokenizers, respectively.

Additionally, Figures S4 and S5 show the relationship between the loss for each continuous property (exact mass, TPSA and Log-P) and the value of its corresponding property. In the case of the exact mass, a clear positive tendency is observed, showing higher losses for extreme values. A similar trend is observed for the TPSA property.

The SMILES that corresponds to the maximum general loss is the PubChem compound CID 102454822. The percentiles of each property value for this compound are 99.545 for the H-bond acceptor count, 11.277 for the H-bond donor count, 0.716 for the rotatable bond count, 99.993 for the exact mass, 99.993 for the TPSA, 99.994 for the heavy atom count and 0.004 for the log-P.

Table 2: Descriptive statistics of the validation loss (mean absolute error).

	Tokenizer	Mean	SD	Min	25%	50%	75%	90%	95%	99%	Max
Loss	Atom-level	2.8090	3.8894	0.0361	0.9133	1.5895	3.0361	5.9854	9.2390	21.1809	202.7497
	SmilesPE	2.1618	3.0629	0.0243	0.7099	1.2829	2.4380	4.6032	6.9208	13.7612	276.9888
H-Bond acceptor	Atom-level	0.0625	0.1752	0.0000	0.0082	0.0184	0.0385	0.1019	0.2638	0.9573	17.7946
	SmilesPE	0.0502	0.1894	0.0000	0.0051	0.0105	0.0187	0.0383	0.1457	1.0050	16.1301
H-Bond donor	Atom-level	0.0226	0.1209	0.0000	0.0005	0.0011	0.0027	0.0123	0.0508	0.9141	5.4683
	SmilesPE	0.0048	0.0383	0.0000	0.0007	0.0016	0.0030	0.0058	0.0098	0.0407	6.0525
Rotatable bond	Atom-level	0.1817	0.3772	0.0000	0.0343	0.0794	0.1758	0.4090	0.7270	1.7737	25.2250
	SmilesPE	0.0949	0.3111	0.0000	0.0099	0.0225	0.0505	0.1775	0.4682	1.2189	23.0964
Exact mass	Atom-level	0.7475	2.6425	0.0000	0.1775	0.3697	0.6768	1.4070	2.8728	5.5871	458.5066
	SmilesPE	0.5678	3.7748	0.0000	0.1169	0.2539	0.4701	0.8185	1.2562	4.3799	547.7298
TPSA[1]	Atom-level	0.5271	1.2293	0.0000	0.1263	0.2768	0.5357	1.0459	1.7687	3.8433	107.6901
	SmilesPE	0.4961	1.3420	0.0000	0.1075	0.2385	0.4821	0.9672	1.5922	4.0695	130.7437
Heavy atom count	Atom-level	0.0299	0.1522	0.0000	0.0056	0.0123	0.0233	0.0461	0.0890	0.3243	35.3366
	SmilesPE	0.0345	0.1986	0.0000	0.0076	0.0165	0.0304	0.0519	0.0776	0.2807	26.9792
Log-P	Atom-level	0.2229	0.2784	0.0000	0.0599	0.1364	0.2816	0.5150	0.7085	1.3265	20.2257
	SmilesPE	0.2219	0.2687	0.0000	0.0631	0.1424	0.2841	0.4988	0.6772	1.2767	15.6543

[1] Topological polar surface area.

Regarding model interpretation, in Figure 1, the embedding of 30,400 randomly selected SMILES generated using the atom-level tokenizer model is observed using the first two principal components. The plots are coloured by six different physicochemical properties. Figure S6 shows the same representation for the SmilesPE tokenizer.

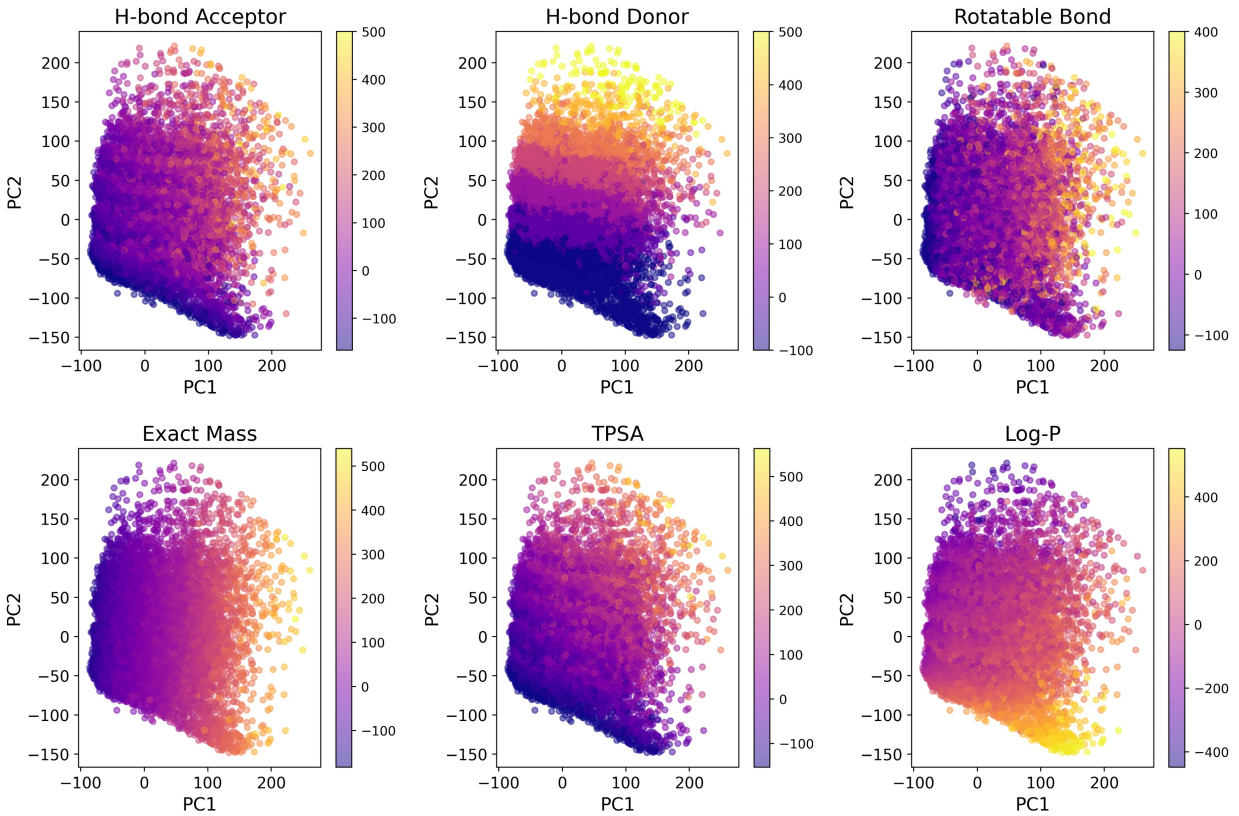

Figure 1: PCAs of the embeddings coloured by six physicochemical properties for the atom-level tokenizer.

In Figure 2, the integrated gradients for two randomly selected molecules in the atom-level case are shown. In Figure S7, the same plots are presented for the SmilesPE tokenizer model. For certain properties, it is evident that the model is scoring the tokens appropriately. For example, for the exact mass, the lighter squares correspond to tokens that contribute to an increase in the exact mass. Similarly, for the heavy atom count, the tokens that represent heavy atoms get higher scores. For the H-bond donor count in D-glucose, which should be 5, only the five oxygen atoms in the hydroxyl groups receive high scores. In contrast, for the H-bond acceptor count, the oxygen atom in the ring is also considered an acceptor, showing high contribution to this property. Regarding the TPSA, only polar atoms are considered, so in the case of D-glucose, only tokens representing oxygen atoms receive high scores. The other molecule analysed is more complex, but similar trends can be observed. In addition,

Figure S7 shows the integrated gradients results for the SmilesPE tokenizer.

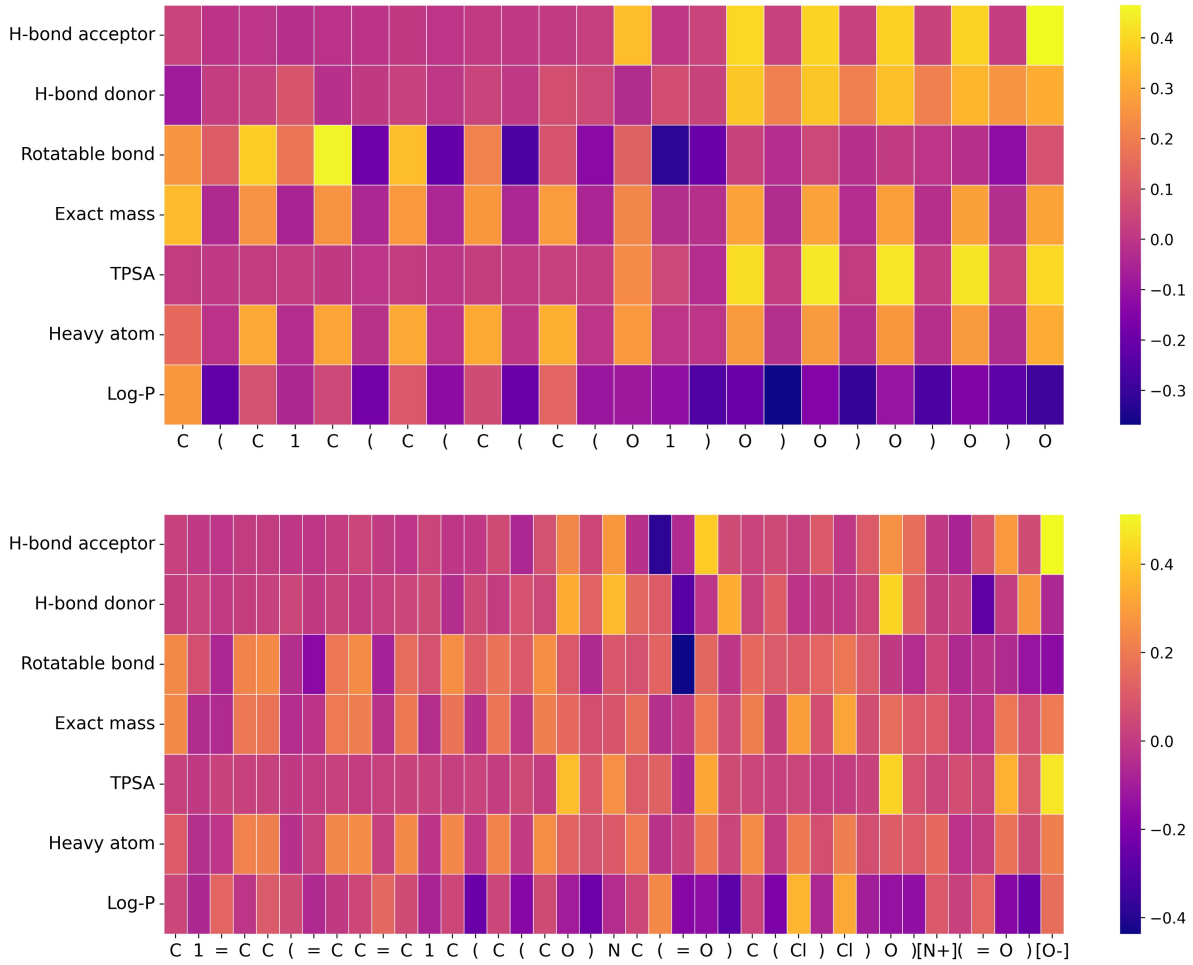

Figure 2: Integrated gradients for two SMILES: D-glucose and 2,2-dichloro-N-[1,3-dihydroxy-1-(4-nitrophenyl)propan-2-yl]acetamide for the atom-level tokenizer

Binding affinity prediction

In Figure 3, the performance of Smile-to-Bert using both atom-level and SmilesPE tokenizers is compared to the Transformer from Morris et al.,[1] trained to translate SMILES strings to IUPAC names. The mean ROC curves obtained using a 10-fold cross-validation strategy and their upper and lower limits created using their standard deviations are shown for trained and untrained embeddings. In Figure 4, the same metrics are shown for two model variations: the

fine-tuned transformer to predict physicochemical properties and the combined embeddings of the Transformer and the large BERT architecture that uses the SmilesPE tokenizer. Both binding affinity datasets AID 652067 and AID 1053197 are used.

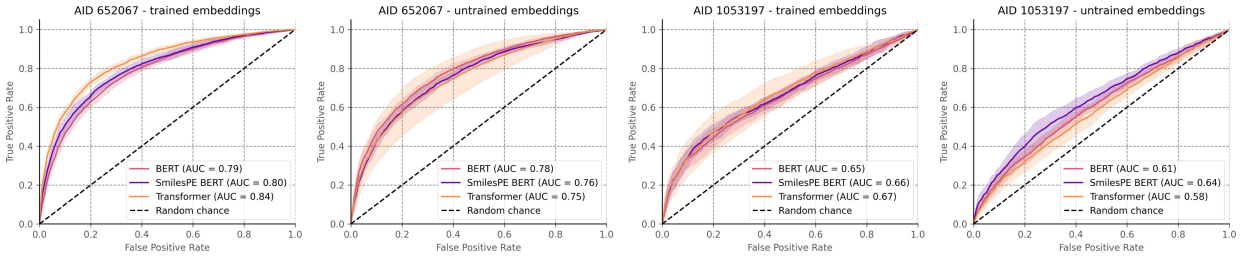

Figure 3: ROC curves for trained and untrained embeddings using three models and both binding affinity datasets.

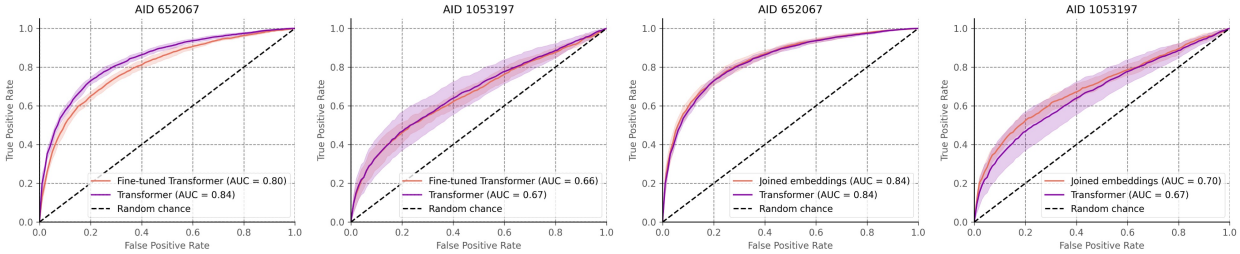

Figure 4: ROC curves for trained embeddings using the pretrained Transformer, a fine-tuned Transformer and joined embeddings for both binding affinity datasets.

Discussion

In this study, we have proven the ability of a BERT architecture to accurately predict physicochemical properties from SMILES, while evaluating two types of tokenizers. Additionally, we have compared the binding affinity prediction capability of our generated embeddings to the ones created by a state-of-the-art Transformer model pre-trained to translate SMILES to IUPAC names in two datasets.

A dashboard application is available to predict seven physicochemical properties from SMILES: H-bond acceptor, H-bond donor, rotatable bond, the exact mass, the TPSA, the heavy atom count and the log-P.

The general validation loss, along with the losses for each property, is presented in Table 2. The mean and standard deviation of each property's loss indicate that the predictions from both BERT architectures are accurate, with exact mass and TPSA having the highest losses. This can be attributed to the fact that these properties exhibit higher values and wider ranges (Table 1). Despite this, the mean exact mass is approximately 363.4 Da, and the average loss for this property is around 0.57 Da for the SmilesPE tokenizer model, highlighting the model's high precision. Although the losses are quite low up to the 99th percentile, the maximum values are very high, indicating that the algorithm is unable to accurately predict a small subset of the SMILES.

A more extensive analysis was performed to elucidate the general characteristics of the SMILES with high losses. On one hand, hexagonal heat maps showing the frequency of compounds in the space between the loss of each continuous property and its corresponding values are presented in Figures S4 and S5 for the atom-level and the SmilesPE tokenizers, respectively. These plots display a clear positive tendency in the case of the exact mass and TPSA, revealing higher losses for larger values, indicating a lower performance for compounds characterised by extreme physicochemical properties values.

On the other hand, the SMILES string corresponding to the maximum general loss was extracted, and it belongs to the PubChem compound CID 102454822. As reported in the Results section, most of the properties of this compound lie in extreme percentiles, contributing to the model's poorer prediction for this specific example.

Additionally, in Table 2, we can see the losses summary for both tokenizers. In general, SmilesPE tokenizer obtains better predictions than the atom-level tokenizer. This suggests that the chemical information contained in the substructures identified by the SmilesPE tokenizer is meaningful for predicting physicochemical properties, among other tasks.

We applied several methods to interpret our models. First, the embeddings were visualised using the first two principal components from a PCA. Figure 1 shows the PCA plots of the embeddings generated by Smile-to-Bert using the atom-level tokenizer, which reveal

a gradient across different properties. Qualitatively, the gradient is more distinct for the H-bond donor count, exact mass, and Log-P. A similar trend can be observed for the embeddings generated by the model using the SmilesPE tokenizer in Figure S6. The second method used for interpretation is the integrated gradients algorithm. Figure 2 displays the importance score for each input token in relation to the property predicted for two molecules. These results help in understanding how the model predicts properties from SMILES. As stated in the Results, the model demonstrates an understanding of how to score the different tokens to predict certain properties, thereby grasping the relationship between atoms and physicochemical properties.

Regarding the prediction of binding affinities, Figure 3 shows the mean ROC curve and its confidence intervals and the mean AUC for our two BERT architectures and the Transformer from Morris et al.[1] Both trained and untrained embeddings are depicted. Overall, the metrics are quite comparable among the three models. However, the Transformer performs better across all datasets, and the SmilesPE tokenizer outperforms the atom-level tokenizer. While the mean AUC achieved by the Transformer is slightly higher for the AID 1053197 dataset, the BERT architectures show better consistency, with narrower confidence intervals. Another interesting observation is that the untrained embeddings for the Transformer perform worse than those for our BERT architectures. This is very likely due to the fact that the Transformer architecture uses character-level tokenization, which makes our input sequences more informative from the beginning.

Several reasons could explain the better performance of the Transformer. On one hand, the embedding vectors in the Transformer case are much longer than the ones obtained in the BERT architectures, therefore, the number of weights in the binding affinity prediction neural network is higher in the Transformer case, which could lead to an enhanced performance. However, lengthier embeddings generated with a large BERT variation do not show an improvement (Figure S8). Another reason could be the type of pre-training task used. While the Transformer is pre-trained to translate SMILES to IUPAC names, probably re-

sulting in embeddings containing general chemical information, in our case, the embeddings obtained are shaped to specifically predict seven physicochemical properties. This could lead to very specific embedding vectors that do not generalise as well as broader ones. In order to test this hypothesis, we fine-tuned the pre-trained Transformer architecture to predict physicochemical properties using our training data. As expected, the two left plots in Figure 4 show a decrease in the mean AUC for the fine-tuned Transformer compared to the original model in both datasets. Therefore, we can hypothesise that a fully supervised pre-training task might not favour the use of the SMILES embeddings for other tasks. Despite this conclusion, we speculate that including information about physicochemical properties, implicitly or explicitly, to the embedding vectors enriches them and favours its ability to generalise well in other tasks. The two right plots in Figure 4 show an improvement in the performance for AID 1053197 when combining the Transformer embeddings together with the large BERT embeddings. This indicates that the integration of physicochemical properties in the embeddings is useful for the prediction of binding affinities, and it might be beneficial for other tasks.

This study has several strengths, including the extensive data used for training and validation, the low losses achieved, and the availability of both the code and a dashboard application for predicting physicochemical properties from SMILES. Additionally, we demonstrate that combining two types of embeddings improves binding affinity prediction performance in one dataset. However, there are some limitations. Although infrequent, some SMILES result in very high prediction losses, indicating that the model struggles with extreme values. This issue could potentially be mitigated through techniques such as data augmentation for rare tokens. Furthermore, while we used a supervised task for pre-training, we show in this work that it generates embeddings that do not generalise as well as those from unsupervised tasks.

Conclusion

In this work, we present Smile-to-Bert, a BERT architecture pre-trained to predict seven physicochemical properties from SMILES, generating contextualised embeddings that incorporate physicochemical properties information. Our model predicts these properties with high accuracy, and using the integrated gradients algorithm, we demonstrate that Smile-to-Bert captures the relationship between atoms and each predicted property. We assessed the utility of these embeddings for binding affinity prediction across two datasets, comparing the performance of two SMILES tokenizers. Combining Smile-to-Bert embeddings with those from a state-of-the-art Transformer improved binding affinity predictions in one dataset. Our findings suggest that a model using the SmilesPE tokenizer, pre-trained with an unsupervised or semi-supervised task, and incorporating physicochemical properties knowledge, could generate richer embeddings that enhance performance in other tasks. A dashboard application is available at `http://147.83.252.32:8050/`, and all code can be accessed at `https://github.com/m-baralt/smile-to-bert`.

Acknowledgement

This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022), by the Grant PID2021-122952OB-I00 funded by AEI 10.13039/501100011033 and by ERDF A way of making Europe; the Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), initiatives of Instituto de Investigación Carlos III (ISCIII); ISCIII (grant AC22/00035); and the CERCA Programme / Generalitat de Catalunya. B2SLab is certified as 2021 SGR 01052. The first author gratefully acknowledges the Universitat Politècnica de Catalunya and Banco Santander for the financial support of her predoctoral grant. Vivian Würf and Josch K.

Pauling were funded by the Bavarian State Ministry of Education and the Arts (StMWK) in the framework of the Bavarian Research Institute for Digital Transformation (bidt, grant LipiTUM).

Supporting Information Available

Details of the Methods and Results.

References

(1) Morris, P.; St. Clair, R.; Hahn, W. E.; Barenholtz, E. Predicting binding from screening assays with transformer network embeddings. *Journal of Chemical Information and Modeling* **2020**, *60*, 4191–4199.

(2) Tran, T.; Ekenna, C. Molecular Descriptors Property Prediction Using Transformer-Based Approach. *International Journal of Molecular Sciences* **2023**, *24*.

(3) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. 2018.

(4) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic acids research* **2023**, *51*, D1373–D1380.

(5) Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine* **2021**, *4*.

(6) Öztürk, H.; Özgür, A.; Schwaller, P.; Laino, T.; Ozkirimli, E. Exploring chemical space using natural language processing methodologies for drug discovery. 2020.

(7) Li, X.; Fourches, D. SMILES Pair Encoding: A Data-Driven Substructure Tokenization Algorithm for Deep Learning. *Journal of Chemical Information and Modeling* **2021**, *61*, 1560–1569.

(8) Schoenmaker, L.; Béquignon, O. J.; Jespers, W.; van Westen, G. J. UnCorrupt SMILES: a novel approach to de novo design. *Journal of Cheminformatics* **2023**, *15*.

(9) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. **2017**,

(10) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*; 2019.

(11) Dong, J.; Wang, N. N.; Liu, K. Y.; Zhu, M. F.; Yun, Y. H.; Zeng, W. B.; Chen, A. F.; Cao, D. S. ChemBCPP: A freely available web server for calculating commonly used physicochemical properties. *Chemometrics and Intelligent Laboratory Systems* **2017**, *171*, 65–73.

(12) Temple Lang, D. RCurl: General Network (HTTP/FTP/...) Client Interface for R. 2024; `https://CRAN.R-project.org/package=RCurl`.

(13) Cao, Y.; Charisi, A.; Cheng, L. C.; Jiang, T.; Girke, T. ChemmineR: A compound mining framework for R. *Bioinformatics* **2008**, *24*, 1733–1734.

(14) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019.

(15) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H.; Murphy, R. C.; Raetz, C. R.; Russell, D. W.; Subramaniam, S. LMSD: LIPID MAPS structure database. *Nucleic Acids Research* **2007**, *35*.

(16) Conroy, M. J.; Andrews, R. M.; Andrews, S.; Cockayne, L.; Dennis, E. A.; Fahy, E.; Gaud, C.; Griffiths, W. J.; Jukes, G.; Kolchin, M.; Mendivelso, K.; Lopez-Clavijo, A. F.;

Ready, C.; Subramaniam, S.; O'Donnell, V. B. LIPID MAPS: update to databases and tools for the lipidomics community. *Nucleic Acids Research* **2024**, *52*, D1677–D1682.

(17) Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019.

(18) Biewald, L. Experiment tracking with weights and biases. *Software available from wandb.com* **2020**, *2*, 233.

(19) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. **2014**,

(20) Sylvain Gugger; Lysandre Debut; Thomas Wolf; Philipp Schmid; Zachary Mueller; Sourab Mangrulkar; Marc Sun; Benjamin Bossan Accelerate: Training and inference at scale made simple, efficient and adaptable. 2022; `https://github.com/huggingface/accelerate`.

(21) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(22) Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; Reblitz-Richardson, O. Captum: A unified and generic model interpretability library for PyTorch. *arxiv* **2020**,

(23) National Center for Biotechnology Information PubChem Database. 2013; `https://pubchem.ncbi.nlm.nih.gov/bioassay/652067`.

(24) National Center for Biotechnology Information PubChem Database. 2014; `https://pubchem.ncbi.nlm.nih.gov/bioassay/1053197`.