PAPER • OPEN ACCESS

SELFormer: molecular representation learning via SELFIES language models

To cite this article: Atakan Yüksel *et al* 2023 *Mach. Learn.: Sci. Technol.* **4** 025035

View the article online for updates and enhancements.

You may also like

- Chemformer: a pre-trained transformer for computational chemistry
  Ross Irwin, Spyridon Dimitriadis, Jiazhen He et al.

- SMILES-X: autonomous molecular compounds characterization for small datasets without descriptors
  Guillaume Lambard and Ekaterina Gracheva

- Curiosity in exploring chemical spaces: intrinsic rewards for molecular reinforcement learning
  Luca A Thiede, Mario Krenn, AkshatKumar Nigam et al.

MACHINE LEARNING
Science and Technology

PAPER

SELFormer: molecular representation learning via SELFIES language models

Atakan Yüksel[1,4], Erva Ulusoy[1,2,4], Atabey Ünlü[1,2,4] and Tunca Doğan[1,2,3,*] [ID]

1 Biological Data Science Lab, Department of Computer Engineering, Hacettepe University, Ankara, Turkey
2 Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Turkey
3 Institute of Informatics, Hacettepe University, Ankara, Turkey
4 Equally contributing co-first authors.
* Author to whom any correspondence should be addressed.

E-mail: tuncadogan@gmail.com

Abstract

Automated computational analysis of the vast chemical space is critical for numerous fields of research such as drug discovery and material science. Representation learning techniques have recently been employed with the primary objective of generating compact and informative numerical expressions of complex data, for efficient usage in subsequent prediction tasks. One approach to efficiently learn molecular representations is processing string-based notations of chemicals via natural language processing algorithms. Majority of the methods proposed so far utilize SMILES notations for this purpose, which is the most extensively used string-based encoding for molecules. However, SMILES is associated with numerous problems related to validity and robustness, which may prevent the model from effectively uncovering the knowledge hidden in the data. In this study, we propose SELFormer, a transformer architecture-based chemical language model (CLM) that utilizes a 100% valid, compact and expressive notation, SELFIES, as input, in order to learn flexible and high-quality molecular representations. SELFormer is pre-trained on two million drug-like compounds and fine-tuned for diverse molecular property prediction tasks. Our performance evaluation has revealed that, SELFormer outperforms all competing methods, including graph learning-based approaches and SMILES-based CLMs, on predicting aqueous solubility of molecules and adverse drug reactions, while producing comparable results for the remaining tasks. We also visualized molecular representations learned by SELFormer via dimensionality reduction, which indicated that even the pre-trained model can discriminate molecules with differing structural properties. We shared SELFormer as a programmatic tool, together with its datasets and pre-trained models at https://github.com/HUBioDataLab/SELFormer. Overall, our research demonstrates the benefit of using the SELFIES notations in the context of chemical language modeling and opens up new possibilities for the design and discovery of novel drug candidates with desired features.

1. Introduction

Traditional methods and techniques of drug discovery are expensive, resource-intensive, and time-consuming, making it impractical to study numerous molecules and identify those that could be developed into new drugs (Vamathevan *et al* 2019). Incorporating novel computational methods into research and development pipelines is crucial, particularly in light of recent events that have highlighted the need for prompt action against sudden disease outbreaks (Basu *et al* 2021). Molecules must be encoded and stored using alphanumeric characters in order to be processed computationally. SMILES (Simplified Molecular-Input Line-Entry System) is a notation for expressing the chemical structure of a molecule using a standard string of linear characters (Wang *et al* 2019). Although SMILES is widely used and has proven

useful in computational chemistry, it is insufficient to fully represent the complexity and properties of molecules. Not every SMILES string corresponds to a valid molecular graph, thereby reducing the representational space (Krenn *et al* 2020). In order to create a better notation for molecules, IUPAC designed InchI, which has a canonical representation that assigns a unique string to each molecule. The InchI system encodes additional information about molecular graphs, including mobile and immobile hydrogens. However, InchI strings are difficult to comprehend, and their complex syntax makes their use in generative modeling challenging (Krenn *et al* 2022). SELFIES is a novel method for representing molecular graphs as character strings that permits the unique identification of any given molecule (Krenn *et al* 2020). SELFIES can be used for multiple purposes, including the construction of molecular fingerprints, the calculation of molecular similarity, and the detection of chemical reactions. SELFIES is human-readable and 100% robust, meaning that every SELFIES string points to a valid molecule, making it more suitable for molecular representation learning than SMILES (Krenn *et al* 2020). SELFIES have been utilized for drug discovery-related tasks in multiple studies (Nigam *et al* 2021, Frey *et al* 2022).

Entities/objects (e.g. molecules) must be represented numerically to be utilized in the artificial learning-based analysis. These representations can be manually crafted (i.e. created empirically using pre-defined rules or descriptions) or automatically extracted in the context of machine learning. Representation learning is concerned with automatically discovering meaningful data representations (Bengio *et al* 2013). These representations are intended to capture the underlying structure and patterns in the data. Some of the primary advantages of representation learning are that; (i) it can help reduce the dimensionality of the data, making it easier to manipulate and interpret (Kopf and Claassen 2021); and (ii) it generates generalized vectors that are potentially reusable in a variety of tasks, such as classification, clustering, and generation (Unsal *et al* 2022). In recent years, numerous examples of representation learning strategies have been published (Chuang *et al* 2020). Self-supervised learning is one example in which representations are learned from large amounts of unlabeled data using pretext tasks such as predicting the next token in an input sequence. These learned representations can then be fine-tuned for specific tasks, such as object recognition or sentiment analysis, for which smaller amounts of labeled data is sufficient (AlBadani *et al* 2022, Wang *et al* 2022a). Overall, representation learning has emerged as an important subfield of artificial learning, as it improves performance and applicability across a broad range of tasks (Ericsson *et al* 2022).

In recent years, transformers have become a widely used architecture for learning self-supervised text representations in the field of natural language processing (NLP) (Vaswani *et al* 2017). Transformers achieved state-of-the-art results on many NLP tasks, including machine translation, text summarization, and question answering. The transformer is based on the concept of self-attention, which allows the model to consider the intrinsic relationships between different words or tokens in a sentence or sequence (Kalyan *et al* 2021). This allows the transformer model to capture long-range dependencies in the data (Tay *et al* 2022). The resulting encodings/embeddings are used as the learned representation of the given sequence (Vaswani *et al* 2017). The transformer-based language modeling approach has a great deal of potential for automatically learning the properties of molecules in large chemical libraries, which can then be used in various steps of drug discovery and development. However, the majority of the proposed chemical language models (CLMs) have a significant shortcoming in that they use SMILES notations to represent molecules. Problems with SMILES-based representations can be listed as; (1) there are a number of ways to write the same molecule in a non-canonical way, which decreases the uniqueness of molecular strings; (2) a valid SMILES string might have invalid chemical properties such as exceeding the natural valency of an atom (Wigh *et al* 2022); (3) SMILES cannot fully capture spatial information; and (4) SMILES alone may not be sufficient to capture molecular characteristics since it lacks syntactic and semantic robustness (Krenn *et al* 2022, Li *et al*, 2022b). These issues suggest a new research direction that may produce CLMs with a greater degree of generalizability. Therefore, we pose the following question: 'Is it possible to obtain a better molecular representation using a chemical language (i.e. a string notation) that is more expressive and suitable for machine learning-based applications than SMILES?'

In this study, we propose a CLM that employs SELFIES representations of molecules in order to obtain their concise, flexible, and meaningful representations for use in a variety of downstream molecular tasks in drug discovery and development. We pre-trained our model, SELFormer, on ~2 million drug-like small molecules in the ChEMBL database, within the framework of self-supervised learning, and evaluated the quality of the output representations in various molecular property prediction-related benchmarks. We provide the method as a programmatic tool, together with trained and ready-to-use models and pre-calculated representation vectors for all ChEMBL molecules. We expect that SELFormer, along with the obtained results, will be useful to computational chemistry researchers in order to facilitate the development of effective solutions for the next generation of drug development.

2. Related work

BERT (Devlin *et al* 2018) is the first major deep learning-based language model that utilizes the transformer architecture. BERT uses two pre-training strategies: masked language modeling (MLM) and next sentence prediction (NSP). In the MLM task, BERT masks 15% of the words in each input sequence by replacing them with a [MASK] token. The model then attempts to predict the original values of these masked words based on their semantic relationships with the non-masked words, which involves adding a classification layer on top of the encoder output and calculating the probability of each word in the vocabulary using the softmax function. The cross-entropy loss function only takes into account the prediction of the masked values, ignoring the prediction of the non-masked values. NSP is a binary classification task in which BERT receives sentence pairs as input and learns to predict whether the second sentence is the one that follows the first sentence in the original input. BERT achieved state-of-the-art results on tasks such as question-answering and language interference. MolBERT (Fabian *et al* 2020) is a CLM designed to learn high-quality and flexible molecular representation using the BERT model and SMILES notations of molecules as the input language. MolBERT is pre-trained with the Guacamole benchmark (Brown *et al* 2019) dataset which is curated from ChEMBL (Gaulton *et al* 2017) consisting of ∼1.6 million molecules. The model was fine-tuned and evaluated on MoleculeNet benchmarking (Wu *et al* 2018) tasks, i.e. BACE, blood–brain barrier penetration (BBBP), and HIV datasets for classification tasks, and ESOL, FreeSolv, and Lipo datasets for regression. MolBERT was tested on three self-supervised tasks: (i) masked language modeling, (ii) predicting the equivalence of two given SMILES inputs where the second SMILES can either be a synonymous permutation of the first one or a SMILES selected from the training data randomly, and (iii) predicting physicochemical properties of the given molecules. The authors found that MolBERT achieved high performance on virtual screening and QSAR benchmarks.

RoBERTa (Liu *et al* 2019) is a BERT-based model that introduces slight improvements over BERT. RoBERTa utilizes dynamic token masking instead of BERT's static token masking, meaning that different tokens are masked in each epoch during training. RoBERTa also removed the NSP task, arguing that it may have a negative effect on the performance by preventing models from learning long-range dependencies. RoBERTa outperformed BERT on the majority of the NLP tasks that were evaluated. ChemBERTa (Chithrananda *et al* 2020) is a model based on RoBERTa and SMILES notations of molecules. The authors of the study investigated how the molecular property prediction performance of transformer-based architectures varies with the size of the pre-training dataset, using masked token prediction as the pre-training objective. The model was pre-trained on 77 million unique molecules curated from PubChem (Kim *et al* 2023) and fine-tuned on MoleculeNet tasks (Wu *et al* 2018). The authors found that ChemBERTa did not outperform the baseline models on most of the downstream tasks, although it did show state-of-the-art performance on toxicity prediction (Tox21). They observed that masked language modeling had a positive impact on downstream task performance. The authors improved the pre-training procedure and achieved higher scores on the same benchmarking tasks in the new version of their model, ChemBERTa-2 (Ahmad *et al* 2022).

Translating the InChI (Handsel *et al* 2021) is a method for predicting the IUPAC name of a chemical from its standard International Chemical Identifier (InChI) using a sequence-to-sequence machine learning model. In lieu of tokenizing the input and output into words or sub-words, this model analyzes the InChI and predicts the IUPAC name character by character. The model achieved a 91% accuracy; however, its results were less precise for inorganic compounds compared to organic molecules.

ChemFormer (Irwin *et al* 2022) is a transformer-based model based on the BART language model. While the BERT model uses masked language modeling for pre-training, BART utilizes denoising of the corrupted sequences by random tokens (Lewis *et al* 2019). The authors pre-trained their models on 100 million SMILES strings curated from the ZINC database. ChemFormer was fine-tuned on downstream tasks such as sequence-to-sequence, regression, synthesis, and retrosynthesis. While sequence-to-sequence tasks are analogous to pre-training; lipophilicity, ESOL, and FreeSolv were used for regression experiments. The synthesis-related task takes two reactants as input and predicts the outcome of the reaction, while the retrosynthesis task was designed as its opposite. The authors reported competitive results against other SMILES and graph-based models.

MolFormer (Ross *et al* 2022) is a CLM that creates molecular representations using SMILES strings of a large molecule dataset curated from the ZINC and PubChem databases. MolFormer was trained with 1.1 billion molecules and then fine-tuned for several classification (e.g. BBBP, Tox21, ClinTox, HIV, BACE, and SIDER) and regression (e.g. QM8, QM9, ESOL, FreeSolv, and lipophilicity) tasks. The model improves on linear attention units (Su *et al* 2021) to speed up the training and lower the memory requirements. In the linear attention unit, MolFormer utilizes rotary positional embeddings instead of absolute positional

embeddings, which was previously proposed in the RoFormer (Su *et al* 2021) study. In return, MolFormer attained improved convergence of the training loss and enhanced stability. MolFormer outperformed all compared models on regression tasks and on half of the classification tasks. The authors concluded that large CLMs perform better than geometric (i.e. graph-based) deep learning models and that, even though no topological information has been fed to the model, it could still distill such knowledge. MolFormer is an important step in creating a large language model; however, MolFormer's large size may limit its accessibility, especially considering the average user/lab. A similar transformer-based method trained via MLM on 1.1 billion molecules, X-MOL (Xue *et al* 2020), also demonstrated the potential of using large-scale unlabelled data for the in silico drug development.

Regression transformer (RT) is a multitask language model that integrates an XLnet backbone and MLM pre-training to leverage the advantages of autoregressive modeling within a bidirectional context (Born and Manica 2023). When predicting properties for small molecules, proteins, and chemical reactions, RT exhibits superior or comparable performance against conventional regression models. Notably, RT extends its capabilities beyond property prediction and excels as a competitive conditional generative model that seamlessly integrates these previously separate tasks without the need for task-specific adjustments.

Born *et al* (2023) presented an interpretable chemical language model (CLM) that combines attention with multiscale convolutions, utilizing data augmentation. They extensively evaluated models built on various molecular representations, including fingerprints, variations of SMILES and SELFIES, as well as graph-based methods. The results showcased the superior performance achieved by incorporating augmentation on the selected molecular representations, leading to the selection of augmented SMILES as the preferred representation for all subsequent analyses. Their model demonstrated competitive predictive performance across various molecular property prediction tasks, particularly the toxicity prediction. The authors highlighted the significance of incorporating diverse architectural enhancements and data augmentation techniques for improved performance and interpretability.

In addition to CLMs, there are methods for learning the representation of molecules that employ molecular graphs. In this approach, atom and bond information are generally encoded as nodes and edges of a molecular graph, respectively. In studies such as SchNet (Schütt *et al* 2017) and MGCN (Lu *et al* 2019), graph representations were utilized to comprehend quantum interactions of small molecules, marking the beginning of graph learning applications to small molecules.

D-MPNN (also known as ChemProp) is a graph-based deep learning model designed to produce molecular embeddings as well as evaluate their properties based on supervised downstream tasks (Yang *et al* 2019). In this study, the authors developed directed message-passing neural networks (D-MPNN) by enhancing the MPNN algorithm (Gilmer *et al* 2017) in order to generate embeddings based on the graph representations of input molecules. The objective of the message-passing algorithm is to determine the properties of the entire graph by aggregating data from connected nodes. ChemProp reported model performance in terms of quantum mechanics, physical chemistry, biophysics, and physiology datasets in the context of downstream classification and regression tasks. On each benchmark, ChemProp outperformed (or competed with) its predecessors. The D-MPNN algorithm is a crucial step in the construction of meaningful molecular graph embeddings. There are models that utilize the modified versions of the MPNN algorithm to achieve similar goals, such as DimeNet (Gasteiger *et al* 2020) and GeomGCL (Li *et al* 2022a). While DimeNet represents molecules based on atomic distances, GeomGCL utilizes both 2D and 3D information. GeomGCL further improved its model with the addition of contrastive learning, which was reported to be beneficial in the case of insufficient labeled data. In their study, Hu *et al* (2019) developed a new strategy for pre-training GNNs at the level of both individual nodes and entire graphs simultaneously, and addressed the difficulties of graph pre-training when task-specific labels are scarce. Other graph-based methods, GEM (Fang *et al* 2022), MolCLR (Wang *et al* 2022b), and Graph MVP-C (Liu *et al* 2021) employ approaches such as graph isomorphism networks (GIN) (Xu *et al* 2018) or contrastive self-supervised learning (Oord *et al* 2018). KPGT (Li *et al* 2022c) benefit from knowledge guided pre-training which leverages molecular descriptors and fingerprints.

There are a few general issues with graph-based models that need to be addressed. Basic GNN methods have limited expressive capabilities and do not perform well in graph isomorphism tests, e.g. the 1-Weisfeiler-Lehman test (Jin *et al* 2017). Higher-order GNNs (Morris *et al* 2019) were introduced to overcome such difficulties. However, increased expressiveness comes at the expense of increased computational requirements. Even though processing molecules in a three-dimensional plane improves performance, it further elevates computational complexity and expense. In addition, large-scale, clean data may not be readily available for 3D modeling. Finally, in some instances, it is reported that these models do not generalize to large molecules (Ross *et al* 2022).

3. Methodology

In SELFormer, we first trained a self-supervised transformer encoder-based CLM to learn high-dimensional embeddings of molecules in our dataset. After that, we fine-tuned the pre-trained model in the context of supervised learning for numerous molecular property prediction tasks by adding a classification or regression head on top of the pre-trained model architecture. Below, we explain each of these steps in detail. The overview of the SELFormer model for pre-training and fine-tuning is shown in figures 1(A) and (B), respectively.

SELFormer is implemented using the Python programming language and the Hugging face Transformers library (Wolf *et al* 2019), which provides a simple and consistent API for building, training, and deploying state-of-the-art machine learning models. Molecules are shared in their SMILES notations in most of the chemistry resources. Therefore, we converted SMILES notations to SELFIES representations using the SELFIES API (Krenn *et al* 2020). To split the fine-tuning datasets into training, and test sets, we utilized the scaffold splitter from the Chemprop library (Yang *et al* 2019).
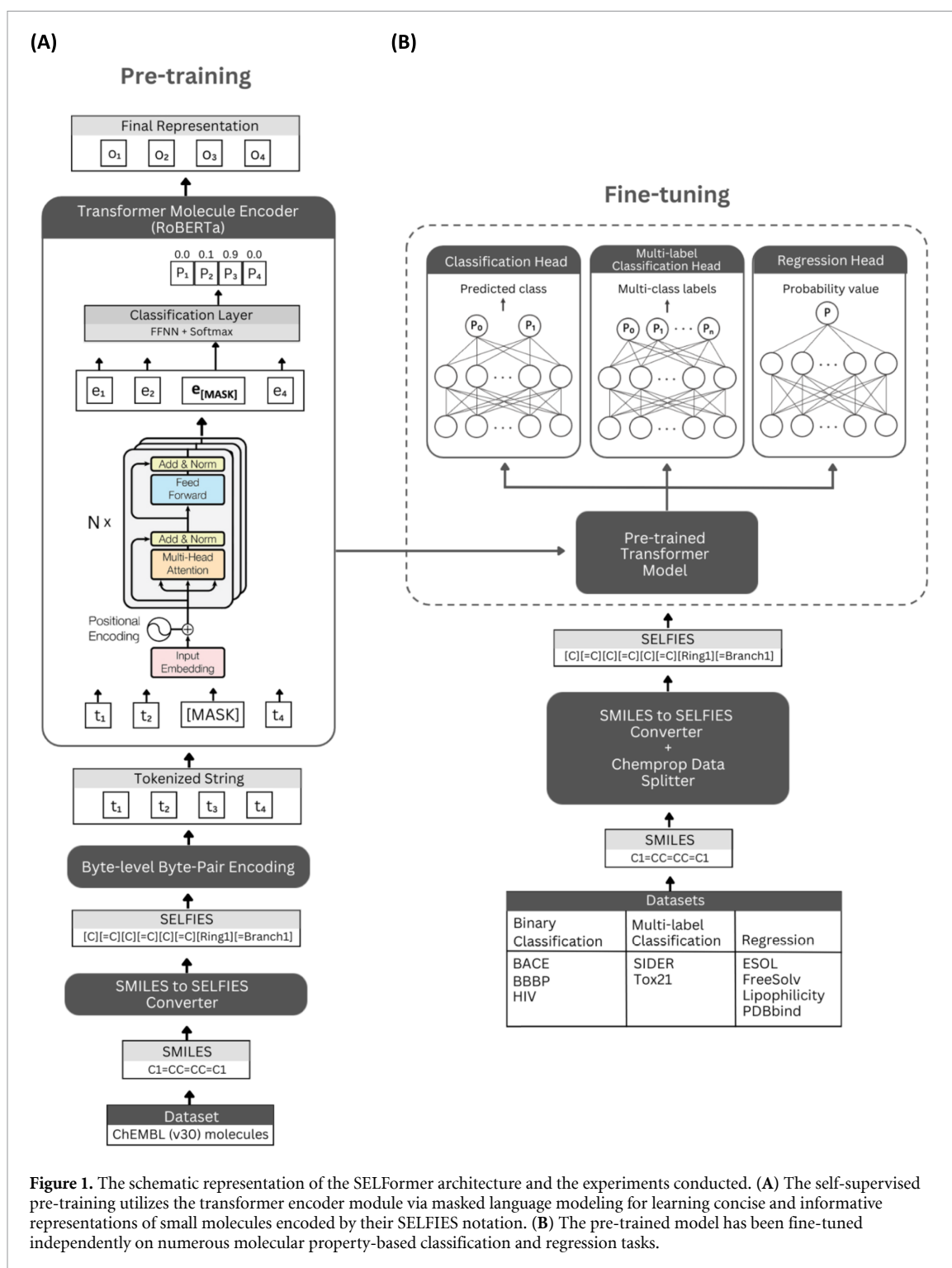
3.1. Pre-training procedure

To pre-train SELFormer (figure 1(A)), we employed the ChEMBL dataset (Gaulton *et al* 2017) (version 30) containing SMILES representations of 2084 725 drug-like bioactive compounds (small molecules). Prior to pre-training, we conducted the SMILES to SELFIES conversion, which resulted in 2084 472 unique molecules in their SELFIES notation. To speed up the conversion process, we parallelized the operation using the Pandaral.lel (https://nalepae.github.io/pandarallel/) Python library. Byte-level byte-pair encoding (BPE) (Radford *et al* 2019) is employed to tokenize the SELFIES notations of the ChEMBL dataset. While defining special tokens, our tokenizer adopts the same approach as RoBERTa (Liu *et al* 2019), enclosing the tokens in angle brackets and using lowercase letters (e.g. <mask>, <unk>, etc.). SELFormer is built on the RoBERTa transformer architecture (Liu *et al* 2019), which utilizes the same architecture as BERT (Devlin *et al* 2018), but with certain modifications that have been found to improve model performance or provide other benefits. One such modification is the use of byte-level BPE (Radford *et al* 2019) for tokenization instead of character-level BPE. Another one is that, RoBERTa is pre-trained exclusively on the MLM objective while disregarding the NSP task.

To optimize our pre-trained models, we conducted a hyperparameter search using a subset of our original training dataset, composed of 100 000 molecules (with a randomized 80–20 train-validation split). Due to computational resource limitations, we followed a sequential approach, resulting in the training of nearly 100 models as opposed to the thousands that would have been trained in a full grid search. In the first hyperparameter search run, we evaluated the impact of the number of attention heads (4, 8, 12), number of hidden layers (4, 8, 12), and learning rate (1×10^{-1}, 1×10^{-3}, 1×10^{-5}, 5×10^{-5}) while keeping batch size and number of epochs constant. We selected the top-performing models based on validation loss values and carried them forward to the second hyperparameter run, in which we varied the batch size (8, 16, 32) and the number of epochs (5, 10, 25) on the best-performing models from the first run. Following the second run, we recognized the interdependence between batch size and learning rate and conducted additional experiments to optimize their relationship. We explored all combinations of learning rate (1×10^{-1}, 1×10^{-3}, 1×10^{-5}, 5×10^{-5}) and batch size (8, 16, 32), while keeping the other hyperparameters constant and aligned with the best-performing model from the second run. Among the ones that performed best, we selected two models to proceed with full-scale pre-training and subsequent fine-tuning. We conducted our experiments on two NVIDIA A5000 GPUs. The total amount of pre-training time for the hyper-parameter optimization tests was around 12 d.

3.2. Fine-tuning

We fine-tuned our models (figure 1(B)) on several molecular property prediction tasks using the MoleculeNet (Wu *et al* 2018) benchmarking datasets of varying sizes including the BBBP, the Side Effect Resource (SIDER), The 'Toxicology in the 21st Century' (Tox21), the ability to inhibit HIV replication (HIV) and binding properties against the human beta-secretase 1 protein (BACE) datasets for classification tasks, and the Free Solvation Database (FreeSolv), aqueous solubility (ESOL), Lipophilicity and the binding affinity prediction (PDBbind) datasets for regression tasks. The details of the selected tasks/datasets are given in table 1.

To be able to fine-tune the SELFormer model for the selected downstream task, we incorporated a classification/regression head consisting of two linear layers on top of the pre-trained model (after the output of the encoder). This module takes the embedding produced by the pre-trained (encoder) model as input, and maps it to the corresponding output, a probability distribution over the target classes for the classification tasks, or a continuous value for the regression tasks.

Figure 1. The schematic representation of the SELFormer architecture and the experiments conducted. (**A**) The self-supervised pre-training utilizes the transformer encoder module via masked language modeling for learning concise and informative representations of small molecules encoded by their SELFIES notation. (**B**) The pre-trained model has been fine-tuned independently on numerous molecular property-based classification and regression tasks.

3.2.1. Classification tasks

The BACE, BBBP, and HIV datasets were used for the binary classification task, while the SIDER and Tox21 datasets were employed for the multi-label classification task. The BACE dataset contains experimental quantitative (i.e. IC50) and qualitative binding properties of a set of inhibitors of human beta-secretase 1 (BACE1) protein (Subramanian *et al* 2016). The dataset comprises 2D structures and binary labels for 1513 compounds. The BBBP dataset provides information on the blood-brain barrier (BBB) permeability properties of over 2,039 compounds (Martins *et al* 2012), a crucial topic for developing drugs that target the central nervous system. The HIV dataset is part of the Drug Therapeutics Program AIDS Antiviral Screen (AIDS antiviral screen data—NCI DTP data—NCI Wiki) and includes over 41 127 compounds categorized as confirmed inactive (CI), confirmed active (CA), or confirmed moderately active (CM) based on their ability to inhibit HIV replication. MoleculeNet has further modified this dataset into binary classification as

Table 1. Information about molecular property prediction tasks that were used in model fine-tuning.

Dataset	Type of the task	Content	# of compounds	Metric
BBBP	Binary classification	Binary labels on blood–brain barrier permeability	2039	ROC-AUC
HIV	Binary classification	Binary labels on the ability to inhibit HIV replication	41 127	ROC-AUC
BACE	Binary classification	Binary labels on human-secretase 1 (BACE1) binding properties	1513	ROC-AUC
SIDER	Multi-label classification	Classification of drug side-effects into 27 system organ classes	1427	ROC-AUC
Tox21	Multi-label classification	Qualitative toxicity measurements on 12 targets	7831	ROC-AUC
ESOL	Regression	Aqueous solubility of common small molecules	1128	RMSE
FreeSolv	Regression	Hydration free energy of small molecules in water	642	RMSE
Lipophilicity	Regression	Experimental octanol/water distribution coefficient of compounds	4200	RMSE
PDBbind	Regression	Binding affinity data for bio-molecular complexes	11 908	RMSE

inactive (CI) and active (CA and CM) compounds. Finally, the SIDER dataset is a collection of marketed drugs and their associated adverse drug reactions (Kuhn *et al* 2016). MoleculeNet has classified the side effects of 1427 approved drugs into 27 categories based on the Medical Dictionary for Regulatory Activities classification system (MedDRA). Additionally, we included the Tox21 dataset, which contains qualitative toxicity measurements for 7831 compounds and can be used for multi-label classification on 12 different targets. To train the models on classification tasks, we utilized cross-entropy (equation (1)) as our loss function. To assess the performance of the models, we calculated the area under receiver operating characteristic curve (ROC), and the area under the precision-recall curve (PRC). On one axis, ROC and PRC curves plot the same metric, True Positive Rate/Recall (equation (2)), while differing in the metric plotted on the other axis, with ROC using false positive rate (equation (3)) and PRC using precision (equation (4))

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i}^{N} \sum_{c=1}^{M} y_{ic} \log\log (\hat{y}_{ic}) \tag{1}$$

where, N is the number of observations, M is the number of classes, y_{ic} is the binary value indicating whether the class label c is the correct classification for the observation i and, $\hat{y}_{ic}$ is the predicted (softmax) probability that the observation i is of class c

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

where TP, TN, FP and FN is the number of true positive, true negative, false positive and false negative predictions, respectively.

3.2.2. Regression tasks
ESOL, FreeSolv, Lipophilicity, and PDBbind datasets were utilized for regression-based evaluation of SELFormer. The ESOL dataset contains aqueous solubility data for 1128 compounds along with their chemical structures (Delaney 2004). FreeSolv, provided by the Free Solvation Database (Mobley and Guthrie 2014) includes experimental and calculated hydration-free energy values of small molecules in water. These values are obtained from molecular dynamics simulation-based alchemical free energy calculations. The Lipophilicity dataset, curated from the ChEMBL database, provides experimental results of octanol/water distribution coefficients for 4200 compounds, which is a critical property for drug molecules, associated to both solubility and membrane permeability. The PDBbind dataset, from the PDBbind database

(Wang *et al* 2004), contains experimentally measured binding affinity values for biomolecular complexes, together with 3D Cartesian coordinates of both ligands and their target proteins derived from experimental measurements. MoleculeNet utilizes the "refined" and "core" subsets of the PDBbind database after careful processing to eliminate data artifacts. Mean squared error loss (equation (5)) was employed as the loss function for fine-tuning the models on regression tasks. To evaluate the performance of the models, we calculated Root Mean Square Error (RMSE) (equation (5)) score

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (5)$$

where; N is the number of observations, y_i is the true value of observation i, and $\hat{y}_i$ is the predicted value of observation i.

For each dataset, we prepared both random and scaffold split datasets. For both, we assumed 80%, 10%, and 10% train, validation, and test split, respectively. We utilized the scaffold splitter from the Chemprop library (Yang *et al* 2019), which divides molecules based on their molecular scaffold, by ensuring that the molecules containing the same scaffold do not appear in more than one split. By partitioning structurally distinct molecules into separate subsets, scaffold splitting provides a more challenging and robust evaluation of a model's learning ability than random splitting (Wu *et al* 2018).

4. Results and discussion

4.1. Pre-training and optimization
To optimize the pre-training procedure we gradually explored hyperparameters by focusing on different combinations, using a random subset of 100 000 molecules from the ChEMBL dataset, as detailed in section 3.1. First, we explored the impact of attention heads, hidden layers, and learning rate via random search, which led to the selection of the top performing hyperparameter values based on validation losses. Second, we examined the effect of batch size and the number of epochs on our top performing models, and found that larger batch sizes and lower number of epochs led to poor performance. We selected two best-performing models with optimal values in terms of the number of attention heads, hidden layers, batch size, and learning rate (table 2). We named the model with fewer trainable parameters SELFormer-Lite, and the larger one, just SELFormer (the default model). Finally, we trained these two models with the full ChEMBL pre-training dataset and calculated their validation loss values as 0.328 and 0.306 for SELFormer-Lite and SELFormer, respectively. Additional experiments conducted on the pre-trained models are presented in section 4.3. The pre-trained models are publicly accessible at https://github.com/HUBioDataLab/SELFormer for further utilization.

4.2. Fine-tuning on molecular property prediction
With the aim of training molecular property prediction models, we fine-tuned the pre-trained SELFormer model on both classification and regression-based tasks, as described in section 3.2.

4.2.1. Classification tasks
With the aim of assessing SELFormer's generalizability and robustness, we calculated its performance on classification-based molecular property prediction tasks using both scaffold and random split datasets. Then, we compared it with the existing CLMs and graph-based molecular representation learning methods. Our evaluation results on the scaffold and random split datasets are presented in tables 3 and 4, respectively.

SELFormer demonstrated high performance especially in two of the selected classification tasks. On SIDER, which is a multi-label classification task concerning drug side effects, SELFormer outperformed all existing approaches, achieving a 10% increase in the overall ROC values compared to the best competitor, MolCLR (tables 3 and 4). On BBBP, a binary classification task for predicting brain-blood barrier permeability, SELFormer outperformed all existing approaches except CLM and KPGT (SELFormer's results are comparable to both CLM and KPGT on this task). It is quite likely that the use of SELFIES representations in our models enhanced the performance in both the SIDER and BBBP tasks as a result of accurately capturing subtle structural differences between molecules (Krenn *et al* 2022), which can have a significant effect on the resulting molecular property value. Furthermore, SELFIES simplifies the representation of stereochemistry compared to the SMILES notation, which in turn reduces the complexity of the input data and facilitates model training. This can be particularly relevant for the prediction tasks related to drug adverse reactions, in which stereochemistry can play an important role. These findings also indicate the advantages of using SELFIES notation for high-throughput drug discovery and molecular design applications.

Table 2. Hyperparameter configurations of the chosen SELFormer models.

	SELFormer-Lite	SELFormer
Batch Size	16	16
Attention Heads	12	4
Hidden Layers	8	12
Learning Rate	5×10^{-5}	5×10^{-5}
Weight Decay	0.01	0.01
Epoch	100	90
# of parameters	58 307 360	86 658 848

Table 3. Classification-based molecular property prediction task (i.e. BACE, BBBP, HIV, Tox21 and SIDER) performance results of SELFormer and competing methods, on scaffold split datasets. Scores are given in terms of the area under receiver operating characteristic curve (ROC) metric (higher is better). Best scores for each task are shown in bold (multiple methods are marked if the difference in-between is less than %1). Scores were not available for the method-task combinations that are denoted with '—'. Results for D-MPNN are taken from Fang *et al* (2022), and results for Hu *et al*, MGCN, GCN, GIN, and SchNet are taken from Wang *et al* (2022b). The HIV result provided for ChemBERTa-2 was adopted from Chithrananda *et al* (2020), which is the previous study of the same authors (i.e. ChemBERTa), as there were no results provided in the ChemBERTa-2 article.

Model type	Model name	BACE	BBBP	HIV	Tox21	SIDER
Text-based	SELFormer (ours)	0.832	0.902	0.681	0.653	**0.745**
	MolBERT	0.866	0.762	0.783	—	—
	ChemBERTa-2	0.799	0.728	0.622	—	—
	CLM	0.861	**0.915**	**0.813**	0.795	0.619
Graph-based	D-MPNN	0.809	0.710	0.771	0.759	0.570
	Hu *et al*	0.859	0.708	0.802	0.787	0.652
	MolCLR	**0.890**	0.736	**0.806**	0.798	0.680
	GraphMVP-C	0.812	0.724	0.770	0.744	0.639
	GEM	0.856	0.724	**0.806**	0.781	0.672
	MGCN	0.734	0.850	0.738	0.707	0.552
	GCN	0.716	0.718	0.740	0.709	0.536
	GIN	0.701	0.658	0.753	0.740	0.573
	SchNet	0.766	0.848	0.702	0.772	0.539
	KPGT	0.855	**0.908**	—	**0.848**	0.649

Table 4. Classification-based molecular property prediction task (i.e. BACE, BBBP, HIV, Tox21 and SIDER) performance results of SELFormer, on random split datasets. Scores are given in terms of the area under receiver operating characteristic curve (ROC) metric (higher is better). Best scores for each task are shown in bold.

Model type	Model name	BACE	BBBP	HIV	Tox21	SIDER
Text-based	SELFormer (Ours)	0.836	0.863	0.714	0.641	**0.745**
	X-MOL	0.872	**0.962**	0.798	—	—
	CLM	—	—	—	**0.858**	0.659
Graph-based	D-MPNN	**0.878**	0.913	**0.816**	0.845	0.646

It is also important to make a direct comparison between SELFormer and ChemBERTa-2, where the latter utilizes the same architecture as SELFormer (i.e. RoBERTa) but SMILES representations of molecules instead of SELFIES. SELFormer outperformed ChemBERTa-2 on BACE and BBBP tasks, and ChemBERTa on the HIV task, with an average of 12% improvement (table 3). As also displayed in table 3, when compared with another text-based method on scaffold split datasets, SELFormer outperformed MolBERT on BBBP with an 18% improvement (MolBERT was not tested on SIDER or Tox21 tasks). MolBERT demonstrated a slight improvement over SELFormer on the BACE task, while on the HIV task, it clearly outperformed SELFormer, possibly due to its incorporation of two supplementary pre-training objectives that consider molecular properties. In the future, extending pre-training with domain-specific objectives can also improve the SELFormer's scores. On the other hand, CLM benefits from the use of SMILES augmentation, which led to better results as observed particularly in the BBBP and HIV tasks.

On scaffold split-based BACE, BBBP, and SIDER tasks, SELFormer mostly dominated graph-based approaches (e.g. with an average of 20% improvement over D-MPNN). The exceptions were for MolCLR, GEM, and KPGT especially considering the tasks of BACE, HIV and Tox21, the superior performance of which can be attributed to their pre-training on larger datasets, such as the entire ZINC (Irwin *et al* 2012)

Table 5. Regression-based molecular property prediction task (i.e. ESOL, FreeSolv, Lipophilicity, and PDBbind) performance results of SELFormer and competing methods, on scaffold split datasets. Scores are given in terms of the root mean squared error (RMSE) metric (lower is better). Best scores for each task are shown in bold. Scores were not available for the method-task combinations that are denoted with '—'.

Model type	Model name	ESOL	FreeSolv	Lipophilicity	PDBbind
Text-based	SELFormer (Ours)	**0.682**	2.797	0.735	1.488
	ChemBERTa-2	—	—	0.986	—
Graph-based	MolCLR	1.110	2.200	0.650	—
	Hu *et al*	1.220	2.830	0.740	—
	D-MPNN	1.050	2.082	0.683	**1.397**
	MGCN	1.270	3.350	1.110	—
	GEM	0.798	**1.877**	0.660	—
	SchNet	1.050	3.220	0.910	—
	KPGT	0.803	2.121	**0.600**	—
	GraphMVP	1.029	—	0.681	—
	GCN	1.430	2.870	0.850	—
	GIN	1.450	2.760	0.850	—

and PubChem (Kim *et al* 2023) databases, and higher number of trainable model parameters. Similarly, there exist a few SMILES-based CLMs, such as MolFormer (Ross *et al* 2022), that were trained on extensive datasets (i.e. 1.1 billion molecules). MolFormer achieved superior performance with a 0.937 ROC value on the BACE task, establishing a new state-of-the-art. MolFormer's performance is not fairly comparable to SELFormer, a model trained on ∼2 million molecules. Nevertheless, on the SIDER task, SELFormer outperformed MolFormer (i.e. the ROC values of 0.745 and 0.690, respectively), again highlighting the benefits of employing the SELFIES notation. On the other hand, SELFormer was outperformed by all graph-based approaches on HIV and Tox21 datasets. This may be explained by the fact that these datasets are relatively larger, which prevented us from exploring a sufficiently large hyperparameter space. Increasing the size of SELFormer and exploring larger hyperparameter spaces by allocating more resources for fine-tuning can improve the model's performance.

Table 4 presents the evaluation results on the randomly split datasets for all classification tasks. The performance of SELFormer on random split datasets is typically higher than that on scaffold split datasets due to the presence of training and test fold molecules that are similar to each other. Wu *et al* (2018) recommend random splitting for Tox21 and SIDER. Due to this, it is especially important to take the random split results into account when analyzing these two tasks. Table 4 reveals that, similar to the scores on scaffold split, SELFormer reported a better performance on the random split SIDER dataset, compared to the other methods.

4.2.2. Regression tasks
SELFormer was also fine-tuned on selected regression-based molecular property prediction tasks in the MoleculeNet benchmark (i.e. ESOL, FreeSolv, Lipophilicity, and PDBbind). These tasks are evaluated based on both scaffold and random split versions of the given datasets. In table 5, the RMSE-based performance of SELFormer is listed with one CLM (i.e. ChemBERTa-2) and multiple graph-based methods, for the scaffold split. SELFormer outperformed all other methods on the ESOL dataset achieving more than a 15% improvement over the RMSE score of the best competitor (i.e. GEM). SELFIES has the ability to successfully express functional groups and aromatic systems, both of which affect the aqueous solubility of molecules (Wang *et al* 2007, Bergström and Larsson 2018). On FreeSolv and Lipophilicity tasks, SELFormer outperformed nearly the half of the graph-based models (i.e. Hu *et al*, GCN, GIN, MGCN, and SchNet) and dominated by the other half (i.e. D-MPNN, GEM, KPGT, GraphMVP-C, and MolCLR). ChemBERTa-2 was the only text-based (i.e. chemical language) model that could be compared with SELFormer. On the Lipophilicity prediction task, SELFormer outperformed ChemBERTa-2.

In table 6, SELFormer was compared with both language models and graph-based methods on the random split datasets (it is advised by Wu *et al* (2018) that ESOL, FreeSolv, and Lipophilicity datasets should be evaluated considering the randomly split versions). While RT, like SELFormer, takes SELFIES as input, MolBERT, Chemformer, and X-MOL were trained using SMILES representations. X-MOL also had the advantage of being trained on a much larger dataset composed of 1.1 billion molecules curated from the ZINC database. Again, SELFormer outperformed all methods on the ESOL dataset while reporting more or less comparable results on the other tasks. Despite the fact that both models utilized SELFIES, SELFormer consistently outperformed RT across all tasks, most likely due to differences in model architecture and

Table 6. Regression-based molecular property prediction task (i.e. ESOL, FreeSolv, Lipophilicity, and PDBbind) performance results of SELFormer and competing methods, on random split datasets. Scores are given in terms of the root mean squared error (RMSE) metric (lower is better). Best scores for each task are shown in bold. Scores were not available for the method-task combinations that are denoted with '—'.

Model type	Model name	ESOL	FreeSolv	Lipophilicity	PDBbind
	SELFormer (Ours)	**0.386**	1.005	0.674	1.437
	MolBERT	0.531	**0.941**	0.561	—
Text-based	Chemformer	0.633	1.230	0.598	—
	X-MOL	0.578	1.108	0.596	—
	RT	0.690	1.030	0.740	
	D-MPNN	0.555	1.075	**0.555**	1.391
Graph-based	SS-GNN	—	—	—	**1.181**
	OnionNET	—	—	—	1.278
	FP-GNN	—	—	—	1.296

Table 7. The ablation study results. Performance of pre-trained and fine-tuned (for 25 and 50 epochs) SELFormer models on classification-based molecular property prediction tasks, using the random split datasets. Results are presented in terms of the area under receiver operating characteristic curve (ROC) and the area under the precision-recall curve (PRC) metrics (higher is better). Best scores for each task are shown in bold.

Model name	BACE		BBBP		HIV		Tox21		SIDER	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
SELFormer-Lite-pretrain	0.682	0.744	0.686	0.914	0.575	0.401	0.500	**0.533**	0.718	0.831
SELFormer-Lite-finetune-25	0.787	0.824	0.826	0.949	0.705	0.463	0.645	0.409	0.708	0.816
SELFormer-Lite-finetune-50	0.773	0.811	0.829	0.950	0.703	0.476	0.650	0.407	0.717	0.819
SELFormer-pretrain	0.699	0.750	0.727	0.925	0.600	0.419	0.523	0.436	**0.723**	**0.831**
SELFormer-finetune-25	**0.796**	**0.830**	0.857	0.958	0.703	0.472	0.641	0.423	0.699	0.806
SELFormer-finetune-50	0.774	0.813	**0.863**	**0.960**	**0.714**	**0.503**	**0.678**	0.489	0.711	0.814

training methodology. SELFormer benefits from large-scale self-supervised pre-training followed by fine-tuning a regression head, while RT relied on training a model based on classification loss (Born and Manica 2023).

On the random split version of the PDBbind dataset, the performance of the SELFormer model was compared against the graph-based models: D-MPNN, SS-GNN (Zhang *et al* 2022), OnionNET (Zheng *et al* 2019), and FP-GNN (Cai *et al* 2022). SELFormer demonstrated comparable performance to these models. Graph-based methods may have an advantage in predicting binding affinities due to their strategy of representing molecules from a geometric perspective, which may better capture their interaction-related properties.
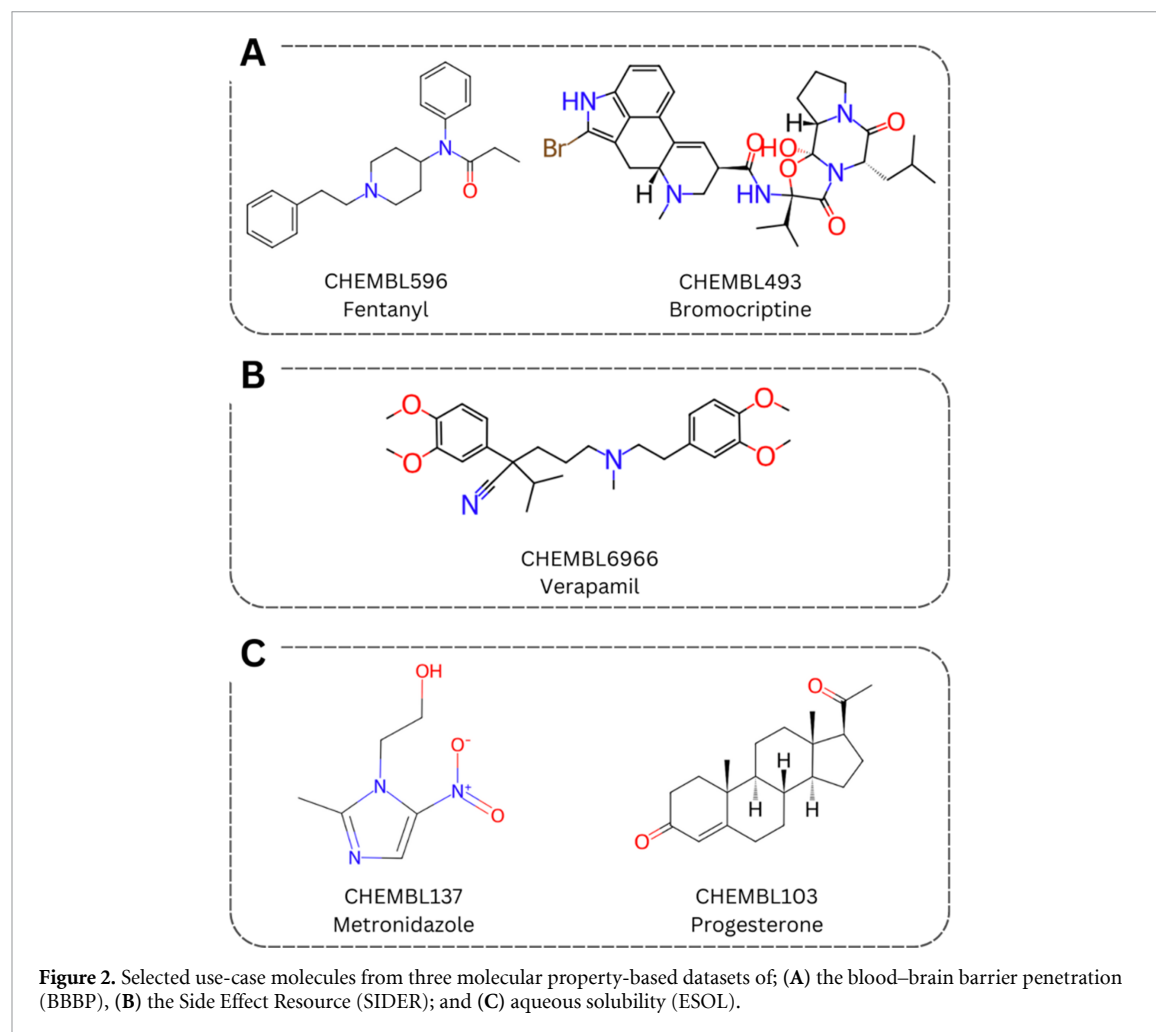
4.3. Ablation study
We conducted an ablation study to investigate the impact of model fine-tuning on downstream prediction performance by comparing the predictive capabilities of two pre-trained models, SELFormer and SELFormer-Lite. We utilized randomly split versions of both the classification- and regression-based molecular property prediction datasets to measure model performances. The training time of the fine-tuned models was restricted by limiting the number of training epochs to 25 and 50, over a fixed set of hyperparameters (i.e., learning rate $= 5 \times 10^{-5}$, train batch size $= 16$, weight decay $= 0$), to yield a basic and fair comparison between different models. In order to calculate the prediction performance of the pre-trained models without fine-tuning, we applied a one-pass learning (i.e., a single epoch training).

The results are presented in tables 7 and 8, for classification and regression tasks, respectively. Our findings revealed that SELFormer consistently outperformed SELFormer-Lite, comparing both their pre-trained and fine-tuned versions against each other. Fine-tuning provided a significant improvement for most of the datasets. Interestingly, we observed that the base (pre-trained) models performed better than their fine-tuned counterparts on the SIDER task, which may be due to the hyperparameters not being optimal for this task. Especially in some of the tasks (e.g., BACE and Lipophilicity) 25-epoch training provided a slightly better performance compared to 50 epochs. Based on these results, we proceeded with fine-tuning experiments on a wider hyperparameter search space using the SELFormer model, as it demonstrated a clear superiority over SELFormer-Lite considering most of the molecular property prediction tasks. The results provided in section 4.2. belong to the abovementioned wider hyperparameter search.

Table 8. The ablation study results. Performance of pre-trained and fine-tuned (for 25 and 50 epochs) SELFormer models on regression-based molecular property prediction tasks, using the random split datasets. Results are presented in terms of the Root Mean Square Error (RMSE) metric (lower is better). Best scores for each task are shown in bold.

Model name	ESOL	FreeSolv	Lipophilicity	PDBbind
	RMSE	RMSE	RMSE	RMSE
SELFormer-Lite-pretrain	1.691	3.124	1.086	1.515
SELFormer-Lite-finetune-25	0.429	1.142	0.725	**1.396**
SELFormer-Lite-finetune-50	**0.380**	1.012	0.698	1.450
SELFormer-pretrain	1.357	3.192	1.021	1.482
SELFormer-finetune-25	0.407	1.054	**0.674**	1.473
SELFormer-finetune-50	0.386	**1.005**	0.711	1.437

Figure 2. Selected use-case molecules from three molecular property-based datasets of; (**A**) the blood–brain barrier penetration (BBBP), (**B**) the Side Effect Resource (SIDER); and (**C**) aqueous solubility (ESOL).

4.4. Use-case analysis

We evaluated the biochemical significance of our findings by analyzing molecular property predictions of selected molecules from downstream analysis datasets, including BBBP, SIDER, and ESOL, in which SELFormer demonstrated a high performance. As use-cases for each task, we chose extensively studied molecules with a high number of publications without considering the SELFormer prediction outcome (figure 2).

    BBBP is a binary classification-based molecular property prediction task comprising the ability of compounds to permeate the BBB. The BBB serves as a selective barrier between the blood circulation and the extracellular fluid of the brain, posing a formidable challenge to the development of drugs that target the central nervous system. We selected two compounds with known effects on the nervous system: fentanyl and bromocriptine. SELformer predicted fentanyl and bromocriptine as positive and negative, respectively, in the BBBP task. Fentanyl, a potent opioid agonist that is indicated for short-term analgesia and is known to cross the BBB (Schaefer *et al* 2017). Additionally, the positive BBB permeability property of fentanyl was also

predicted with high probability (0.9901) by admetSAR (https://go.drugbank.com/drugs/DB00813), an open-source database that collects and curates absorption, distribution, metabolism, excretion, and toxicity (ADMET)-associated property data from published literature (Cheng *et al* 2012). On the contrary, bromocriptine, a dopaminergic receptor agonist used to treat galactorrhea due to hyperprolactinemia and early Parkinson's disease, was predicted to have negative BBB permeability by admetSAR with a probability of 0.9845 (https://go.drugbank.com/drugs/DB01200), which was also predicted as negative by SELFormer.

SIDER is a multi-label classification task of grouping drug side-effects into 27 classes based on the affected system organ. We assessed the performance of SELFormer on this task by examining verapamil, a well-studied calcium channel blocker that is prescribed for angina, arrhythmia, and hypertension (https://go.drugbank.com/drugs/DB00661). The results demonstrate that SELFormer predicted several system organ classes associated with verapamil's side-effects, including infections, infestations, respiratory and gastrointestinal disorders. The SIDER database reports that verapamil commonly causes infections, constipation, and upper respiratory tract infections, with frequencies ranging from 2.4% to %12.1, which are all related to system organ classes predicted by SELFormer (http://sideeffects.embl.de/drugs/2520). Cardiac failure and pulmonary oedema are other reported frequent side-effects related to vascular, respiratory, thoracic and mediastinal disorders which were also correctly predicted by SELFormer. One of the classes predicted by SELFormer for verapamil was 'metabolism and nutrition disorders', which was evaluated to be a false-positive prediction based on the annotations of this molecule in SIDER. Calcium channel blockers have been reported to inhibit insulin release, and some studies have shown reversible hyperglycemia in patients with maturity onset diabetes mellitus treated with these drugs. Also, short-term administration of calcium channel blockers to individuals with normal glucose tolerance was reported to result in a significant glucose intolerance (Russell 1988). These findings suggest that verapamil may have metabolic effects beyond its known side-effects, although additional studies are needed to confirm their clinical relevance and mechanisms. The ability of SELFormer to predict such potential side-effects across multiple system organ classes demonstrates the model's high generalization ability for drug side-effect prediction.

The ESOL dataset is a regression task that aims to predict the solubility of compounds in water based on their chemical structures. Aqueous solubility is a key property for drug discovery and molecular design as it affects drug absorption, distribution, metabolism, and excretion in the body. We selected two compounds, metronidazole and progesterone, to evaluate SELFormer on ESOL. Metronidazole is an antibiotic from the nitroimidazole class, commonly used to treat gastrointestinal infections, trichomoniasis, giardiasis, and amoebiasis (Dingsdag *et al* 2018, Hernández Ceruelos *et al* 2019). The measured log solubility of metronidazole reported in the ESOL dataset was 10.321 g l^{-1} at 25 °C (Delaney 2004), while SELFormer predicted a value of 6.659 g l^{-1}, which represented a 35% deviation from the measured value. Progesterone is a vital hormone for endometrial receptivity, embryo implantation, and successful pregnancy, and is also used in contraceptive preparations to prevent ovulation and fertilization and in other formulations to support pregnancy (Cooper *et al* 2022). For progesterone, the measured log solubility in the ESOL dataset was 0.012 g l^{-1}, and SELFormer predicted a value of 0.028 g l^{-1}. Although SELFormer did not achieve particularly high performance on the use-cases we evaluated for the ESOL dataset, it is worth noting that aqueous solubility calculations are not always precise, leading to varying reported values across different databases. For instance, the reported aqueous solubility of metronidazole in DrugBank is 5.92 g l^{-1} (https://go.drugbank.com/drugs/DB00916), which is more consistent with SELFormer's prediction than with the ground truth value in the ESOL dataset.

4.5. Visualization of the embedding space

Dimensionality reduction techniques are frequently employed to visually explore the distribution of points in a dataset, based on their features, with the aim of conducting qualitative evaluations. Here, we investigated the molecular representations produced by SELFormer via uniform manifold approximation and projection (UMAP), which is a general purpose manifold-learning-based dimensionality reduction method (McInnes *et al* 2018).

Due to the simplicity of visual examination and comparison between two groups, we created UMAP embeddings for the molecules in our binary classification-based molecular property prediction tasks (i.e. BACE, BBBP, and HIV). For each task, we constructed one embedding using the representations obtained from the pre-trained model, and another one using the representations obtained from the fine-tuned model. The same set of molecules were employed for the pre-trained and fine-tuned embeddings of the same task. The embeddings are given in figure 3.

The molecules of the BACE task were visualized using the whole dataset since it is relatively small (i.e. composed of 1513 data points). The two groups in figure 3(A) represent the active and inactive molecules against the human beta-secretase 1 (BACE1) protein. In both pre-trained and fine-tuned model embeddings, active and inactive molecules are separated from each other, although some inactive molecules
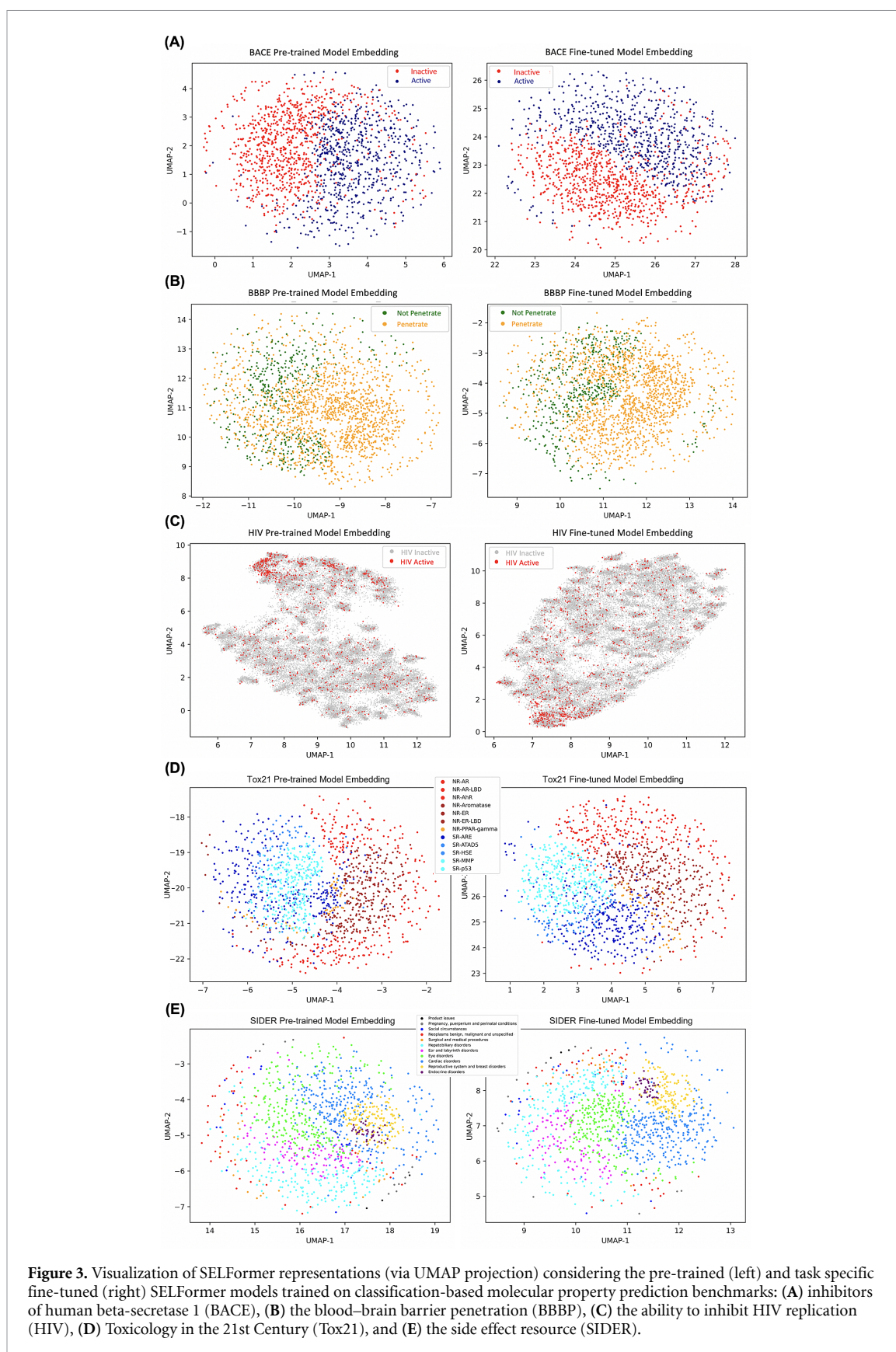
Figure 3. Visualization of SELFormer representations (via UMAP projection) considering the pre-trained (left) and task specific fine-tuned (right) SELFormer models trained on classification-based molecular property prediction benchmarks: (**A**) inhibitors of human beta-secretase 1 (BACE), (**B**) the blood–brain barrier penetration (BBBP), (**C**) the ability to inhibit HIV replication (HIV), (**D**) Toxicology in the 21st Century (Tox21), and (**E**) the side effect resource (SIDER).

appear to be scattered throughout the area occupied by active molecules; however, in the fine-tuned model, two groups are slightly more linearly separable, which indicates that the fine-tuning process improved the representation power of SELFormer in the BACE task.

The BBBP dataset was embedded in figure 3(B), with the two groups representing molecules that can penetrate the BBB and the ones that cannot. The dataset in the BBBP task is not well balanced between

classes, as there is a higher number of molecules that can penetrate the brain-blood barrier. In pre-trained model embeddings, two classes are roughly separated into the two sides of the plane. Also, fine-tuning slightly improved the clustering of the non-penetrable molecules by bringing their two clusters closer to each other.

The HIV dataset contains ~41 000 molecules making it the largest classification dataset SELFormer models were fine-tuned on. This dataset is extremely unbalanced with only ~1400 positive (active) and ~39 600 negative (inactive) data points. For HIV active molecules, both pre-trained and fine-tuned model embeddings formed numerous clusters, a few of which were large and contained most of the members of the group (figure 3(C)). This result shows that the models were able to capture a certain amount of structural similarity between HIV active molecules. There were small clusters of inactive molecules in the embeddings as well; however, they did not cluster well, probably due to their high number and diversity. As an overall evaluation, UMAP embeddings indicate SELFormer was able to learn the distinction between small molecules across multiple types of molecular properties, even only with pre-training. Additionally, fine-tuning causes a slight improvement. It is also important to note that, assessing the magnitude of improvement over 2D visualizations may not be ideal, since the dimension size is not sufficient to fully express the features learned by the model.

For the embeddings of molecules in the Tox21 and SIDER datasets, preprocessing steps were employed, mainly to address the inherent challenges of representing multi-class and multi-label data on the 2D plane. For Tox21, the molecules with multiple positive labels, or without any positive labels, were eliminated, resulting in a dataset comprising 1393 molecules. These molecules possess single positive labels distributed across 12 distinct Tox21 classes. The Tox21 dataset consists of two main groups/panels: the nuclear receptor panel (NR) and the Stress Response panel (SR). In the UMAP plots, we colored molecules according to their Tox21 classes (i.e. the sub-groups of SR and NR) in a way to improve visual perception. As observed in figure 3(D) (left), using the pre-trained SELFormer model, we were able to distinguish between the SR and NR panels. The SR group molecules, represented by a blue palette, are positioned on the left side of the UMAP plot, while the NR group molecules, represented by a red palette, are embedded on the right side. Furthermore, by fine-tuning the SELFormer model on Tox21 (figure 3(D), right side), it was possible to achieve a greater discrimination between the sub-groups of SR and NR.

In SIDER, there is a high degree of overlap between classes, due to the fact that a significant proportion of molecules exhibited multiple positive labels. To address this problem, we utilized a simple inter-class pairwise similarity function and calculated the ratio of shared molecules between each pair of classes in a manner similar to the Jaccard index. Using this function, we eliminated all classes (and their associated molecules) that have a similarity score of 0.8 or greater against all other classes. In addition, among the remaining molecules, those with more than three class assignments were eliminated. These criteria were adopted to ensure a substantial level of dissimilarity between selected samples. In the end, we obtained a total of 1170 positive assignments/labels across 11 side effect classes. The UMAP embedding in the left side of figure 3(E) shows the pretrained model embedding for SIDER. Although there are overlaps between classes, the pretrained model can be considered successful in differentiating side effect classes in general. On the other hand, the fine-tuned SELFormer model (figure 3(E), right side) exhibited an enhanced ability to distinguish side effect classes. A few examples of improvements over the pre-trained model can be given as the separation between the molecules that are associated with; (i) the reproductive system and breast disorders (gold color) and cardiac disorders (light blue color), and (ii) eye disorders (green color) and cardiac disorders (light blue color). Notably, for both Tox21 and SIDER, obtaining visually striking clusters from the 2D embeddings of molecules remained challenging, despite the aforementioned data preprocessing steps, due to the multi-class and multi-label characteristics of these tasks.

To evaluate the pre-trained SELFormer model from a drug discovery perspective, we generated a UMAP projection and 2D visualization (figure 4) of a large-scale molecule dataset based on compound-target interactions (i.e. bioactivity data) obtained from the ChEMBL database v29. The specific aim behind this analysis was to observe whether it is possible for SELFormer to distinguish molecules based on their functional/interaction related properties (i.e. grouping them based on their target proteins' families), which is a highly difficult task due to the fact that molecules with dissimilar structures can interact with different members of a large protein family, and can even bind to the same region on a particular protein in some cases. The dataset we used for this experiment contains small molecule ligands (inhibitors) of proteins from five different protein families which are transferases, proteases, oxidoreductases, membrane-receptors, and ion-channels. While gathering this dataset, we only included the molecules with an experimentally measured bioactivity value of pChEMBL > 6 (i.e. IC50 < 1uM) in single target-based binding assays. For the UMAP embedding, we randomly selected 10 000 inhibitor molecules from each protein family and eliminated molecules that are recorded to be bioactive against multiple families. Only the pre-trained SELFormer model's representations were used in this UMAP analysis, as we did not have a model fine-tuned for this task.
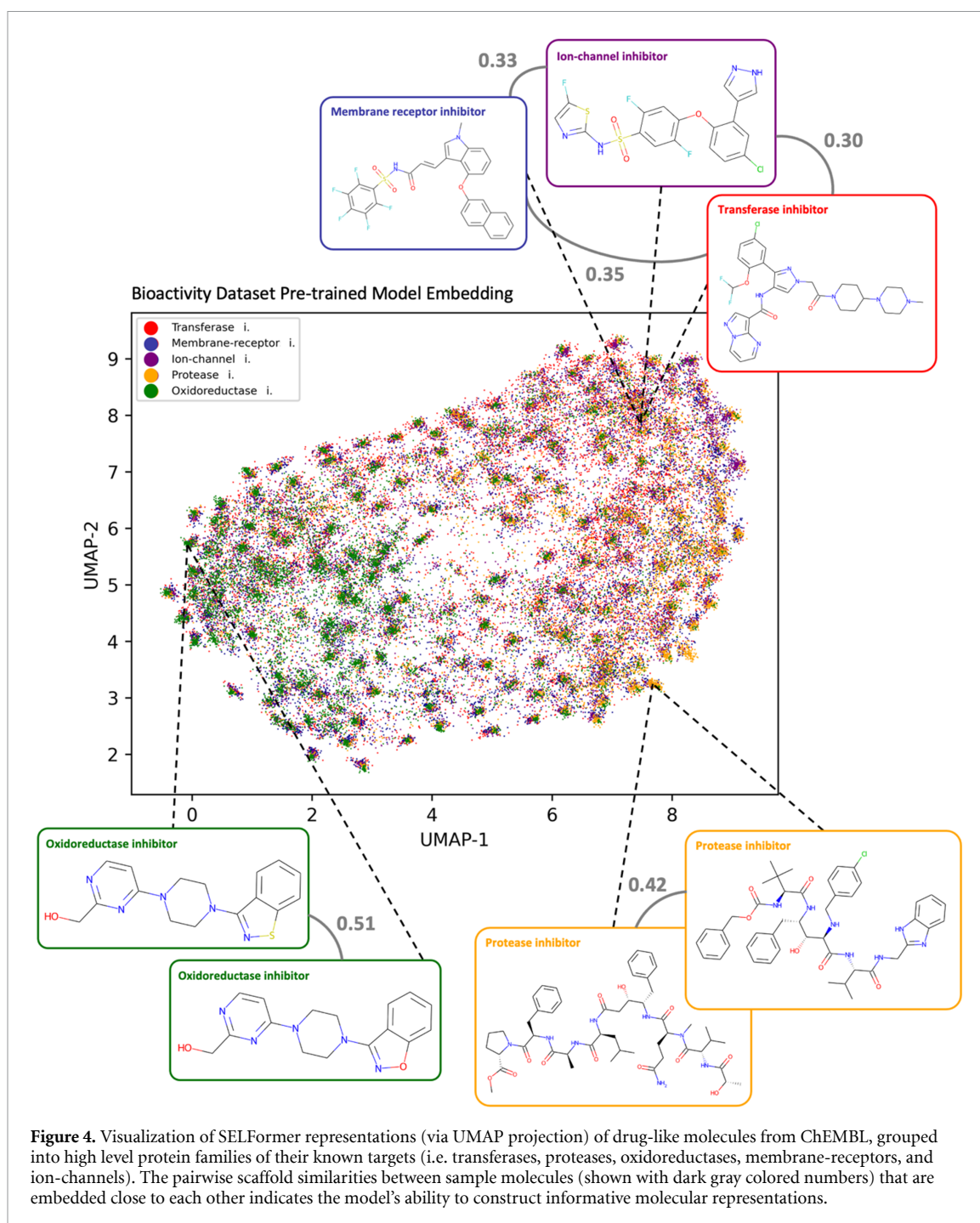
Figure 4. Visualization of SELFormer representations (via UMAP projection) of drug-like molecules from ChEMBL, grouped into high level protein families of their known targets (i.e. transferases, proteases, oxidoreductases, membrane-receptors, and ion-channels). The pairwise scaffold similarities between sample molecules (shown with dark gray colored numbers) that are embedded close to each other indicates the model's ability to construct informative molecular representations.

In figure 4, it can be seen that molecules formed numerous small clusters, some of which are homogeneous and some are heterogeneous considering the protein families of their targets. This is most probably due to apparent structural similarities between molecules. Oxidoreductase inhibitors are clustered into several groups and located abundantly at the left hand side of the plane. Whereas, most of the protease inhibitors are located at the right-bottom side of the plot, again in several small clusters. Transferases group contains kinase inhibitors which are diverse in both structure and function, and this is reflected as scattered red dots all over the embedding plane. A similar situation applies for the membrane receptor inhibitors.

We selected sample molecules from two highly homogenous clusters (mostly composed of oxidoreductase and protease inhibitors, respectively) and one heterogeneous cluster (composed of inhibitors from all protein families) for further analysis. We calculated pairwise Bemis–Murcko scaffold-based (Bemis and Murcko 1996) Tanimoto similarities between selected molecules that are close to each other in the UMAP embedding space. Selected oxidoreductase inhibitors had a scaffold similarity of 0.51, whereas the similarity between selected protease inhibitors was 0.42. Moreover, three molecules from the heterogeneous cluster had a mean scaffold similarity of 0.33. Their Kekule representations also indicate the structural

resemblance (figure 4). We also calculated inter-cluster similarities for the same clusters, which are far away from each other on the embedding plane. The mean pairwise scaffold similarities between the members of cluster pairs was found as; oxidoreductase vs. protease: 0.18 ± 0.04, oxidoreductase vs. heterogeneous: 0.22 ± 0.06, and protease vs. heterogeneous: 0.25 ± 0.08. These results demonstrate that SELFormer can accurately capture structural relationships, as structurally similar molecules are placed adjacently and dissimilar molecules are embedded to distant points (in different clusters) on the 2D plane. We observed that even the heterogeneous cluster have structurally similar members and the heterogeneity mostly originates from the low correlation between structural and target interaction-related similarities. One probable explanation for the low correlation is that most of the drug candidate molecules have been screened against only one target, or a few targets from the same family. As a result, it is possible that they interact with the members of the other protein family as well (of course, this is not known presently), which would change the grouping patterns in our embedding. Overall, distinct target protein family-based clusters could not be observed mainly due to missing data and the high difficulty of this task. Nevertheless, SELFormer could capture a certain amount of functional information mostly in relation to structural similarities.

5. Impact and future directions

In this study, we introduced SELFormer, a new method for learning the representation of the chemical space that leverages the SELFIES notations of molecules and the transformer architecture. Our evaluation of SELFormer's performance on classification and regression-based molecular property prediction tasks revealed that it performs well and surpasses existing approaches on multiple tasks, such as predicting drug side effects, brain-blood barrier permeability, and aqueous solubility. However, larger datasets such as HIV and Tox21 presented challenges in our study mainly due to the requirement to explore large hyperparameter spaces. Comparison of SELFormer to existing language modeling-based approaches demonstrates the advantages of SELFIES notation over SMILES in molecular property prediction, including capturing diverse and complex molecular structures, simplifying stereochemistry representation, and thereby improving model accuracies.

The results of the ablation study suggest that careful fine-tuning of models can significantly improve performance in molecular property prediction. We further visualized the performance of the pre-trained and fine-tuned SELFormer models on binary classification-based tasks by generating UMAP embeddings. The results showed that SELFormer was able to learn the distinction between small molecules across different molecular properties, even only with pre-training. Fine-tuning the pre-trained models yielded a slight improvement in 2D embeddings. An additional UMAP projection and visualization for a dataset based on compound-target interactions demonstrated that SELFormer was able to capture a certain amount of functional (bioactivity related) information, although distinct target protein family-based inhibitor molecule clusters could not be observed.

Considering the size of our pre-training dataset and the total number of trainable parameters in the model (i.e. 2 million molecules and ∼86 million parameters, respectively), it is possible to state that SELFormer is not a 'large model'. Nevertheless, it performed well on the respective supervised prediction tasks. We suggest that pre-training the SELFormer model with larger molecule datasets, such as the entire ZINC database (Irwin *et al* 2012) and enlarging the model by increasing the number of learnable parameters could lead to even better results, although additional experiments are needed to confirm this.

An interesting direction for molecular representation learning is the development of multi-modal models in which the structural data is processed together with other types of molecular information such as natural language-based text (from articles and other types of documents), and functional annotations (e.g. bio-interactions against targets, ontological associations, indications, etc.). Recent studies have integrated SMILES notations of molecules into: (i) textual descriptions (Liu *et al* 2022), and (ii) clinical semantic information (Wen *et al* 2023) to generate multi-modal molecule representations. To further enhance molecular understanding and representation learning within these models, an interesting experiment would involve the utilization of SELFIES in place of SMILES. Furthermore, biomedical knowledge graphs, such as the ones provided by the CROssBAR system (Doğan *et al* 2021), in which target protein interactions, signaling/metabolic pathway associations, and phenotype (Doğan 2018) and disease indications of drugs and bioactive compounds are represented, would be an intriguing type of data structure to leverage in this regard.

Overall, the SELFormer model exhibited high potential in molecular representation learning, which can translate well into improvements in high-throughput drug discovery and molecular design. With further exploration and optimization, SELFormer, or SELFIES-based representation models in general, could serve as a powerful and persistent alternative to SMILES-based CLMs for encoding and analyzing complex molecular structures.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/HUBioDataLab/SELFormer.

Acknowledgments

ORCID iD

Tunca Doğan ⓘ https://orcid.org/0000-0002-1298-9763

References

Ahmad W, Simon E, Chithrananda S, Grand G and Ramsundar B 2022 ChemBERTa-2: towards chemical foundation models (arXiv:2209.01712)

AlBadani B, Shi R and Dong J 2022 A novel machine learning approach for sentiment analysis on twitter incorporating the universal language model fine-tuning and SVM *Appl. Syst. Innov.* **5** 13

Basu S, Ramaiah S and Anbarasu A 2021 In-silico strategies to combat COVID-19: a comprehensive review *Biotechnol. Genet. Eng. Rev.* **37** 64–81

Bemis G W and Murcko M A 1996 The properties of known drugs. 1. Molecular frameworks *J. Med. Chem.* **39** 2887–93

Bengio Y, Courville A and Vincent P 2013 Representation learning: a review and new perspectives *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1798–828

Bergström C A and Larsson P 2018 Computational prediction of drug solubility in water-based systems: qualitative and quantitative approaches used in the current drug discovery and development setting *Int. J. Pharm.* **540** 185–93

Born J and Manica M 2023 Regression Transformer enables concurrent sequence regression and generation for molecular language modelling *Nat. Mach. Intell.* **5** 432–44

Born J, Markert G, Janakarajan N, Kimber T B, Volkamer A, Martínez M R and Manica M 2023 Chemical representation learning for toxicity prediction *Digit. Discovery* (https://doi.org/10.1039/D2DD00099G)

Brown N, Fiscato M, Segler M H and Vaucher A C 2019 GuacaMol: benchmarking models for de novo molecular design *J. Chem. Inf. Model.* **59** 1096–108

Cai H, Zhang H, Zhao D, Wu J and Wang L 2022 FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction *Brief. Bioinformatics* **23** bbac408

Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G, Lee P W and Tang Y 2012 admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties *J. Chem. Inf. Model.* **52** 3099–105

Chithrananda S, Grand G and Ramsundar B 2020 ChemBERTa: large-scale self-supervised pretraining for molecular property prediction (arXiv:2010.09885)

Chuang K V, Gunsalus L M and Keiser M J 2020 Learning molecular representations for medicinal chemistry: miniperspective *J. Med. Chem.* **63** 8705–22

Cooper D B, Patel P and Mahdy H 2022 Oral contraceptive pills (StatPearls—NCBI Bookshelf) (available at: www.ncbi.nlm.nih.gov/books/NBK430882/) (Accessed 30 March 2023)

Delaney J S 2004 ESOL: estimating aqueous solubility directly from molecular structure *J. Chem. Inf. Comput. Sci.* **44** 1000–5

Devlin J, Chang M W, Lee K and Toutanova K 2018 BERT: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805)

Dingsdag S A and Hunter N 2018 Metronidazole: an update on metabolism, structure-cytotoxicity and resistance mechanisms *J. Antimicrob. Chemother.* **73** 265–79

Doğan T 2018 HPO2GO: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences *PeerJ* **6** e5298

Doğan T *et al* 2021 CROssBAR: comprehensive resource of biomedical relations with knowledge graph representations *Nucleic Acids Res.* **49** e96

Ericsson L, Gouk H, Loy C C and Hospedales T M 2022 Self-supervised representation learning: introduction, advances, and challenges *IEEE Signal Process. Mag.* **39** 42–62

Fabian B, Edlich T, Gaspar H, Segler M, Meyers J, Fiscato M and Ahmed M 2020 Molecular representation learning with language models and domain-relevant auxiliary tasks (arXiv: 2011.13230)

Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, Wang F, Wu H and Wang H 2022 Geometry-enhanced molecular representation learning for property prediction *Nat. Mach. Intell.* **4** 127–34

Frey N, Soklaski R, Axelrod S, Samsi S, Gomez-Bombarelli R, Coley C and Gadepally V 2022 Neural scaling of deep chemical models *ChemRxiv* (https://doi.org/10.26434/chemrxiv-2022-3s512) (Accessed 09 June 2023)

Gasteiger J, Groß J and Günnemann S 2020 Directional message passing for molecular graphs (arXiv:2003.03123)

Gaulton A *et al* 2017 The ChEMBL database in 2017 *Nucleic Acids Res.* **45** D945–54

Gilmer J, Schoenholz S S, Riley P F, Vinyals O and Dahl G E 2017 Neural message passing for quantum chemistry *34th Int. Conf. on Machine Learning* (*Sydney, Australia, 11–15 August 2017*) (PMLR) pp 1263–72 (available at: https://proceedings.mlr.press/v70/gilmer17a)

Handsel J, Matthews B, Knight N J and Coles S J 2021 Translating the InChI: adapting neural machine translation to predict IUPAC names from a chemical identifier *J. Cheminformatics* **13** 1–11

Hernández Ceruelos A, Romero-Quezada L C, Ruvalcaba Ledezma J C and López Contreras L 2019 Therapeutic uses of metronidazole and its side effects: an update *Eur. Rev. Med. Pharmacol. Sci.* **23** 397–401

Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V and Leskovec J 2019 Strategies for pre-training graph neural networks (arXiv:1905.12265)

Irwin J J, Sterling T, Mysinger M M, Bolstad E S and Coleman R G 2012 ZINC: a free tool to discover chemistry for biology *J. Chem. Inf. Model.* **52** 1757–68

Irwin R, Dimitriadis S, He J and Bjerrum E J 2022 Chemformer: a pre-trained transformer for computational chemistry *Mach. Learn.: Sci. Technol.* **3** 015022

Jin W, Coley C, Barzilay R and Jaakkola T 2017 Predicting organic reaction outcomes with Weisfeiler-Lehman network *Advances in Neural Information Processing Systems* (*Long Beach, California, USA, 4–9 December 2017*) vol 30 (available at: https://proceedings.neurips.cc/paper/2017/hash/ced556cd9f9c0c8315cfbe0744a3baf0-Abstract.html)

Kalyan K S, Rajasekharan A and Sangeetha S 2021 Ammus: a survey of transformer-based pretrained models in natural language processing (arXiv:2108.05542)

Kim S *et al* 2023 PubChem 2023 update *Nucleic Acids Res.* **51** D1373–80

Kopf A and Claassen M 2021 Latent representation learning in biology and translational medicine *Patterns* **2** 100198

Krenn M *et al* 2022 SELFIES and the future of molecular string representations *Patterns* **3** 100588

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

Kuhn M, Letunic I, Jensen L J and Bork P 2016 The SIDER database of drugs and side effects *Nucleic Acids Res.* **44** D1075–9

Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V and Zettlemoyer L 2019 Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (arXiv:1910.13461)

Li H, Zhao D and Zeng J 2022c Kpgt: knowledge-guided pre-training of graph transformer for molecular property prediction (arXiv:2206.03364)

Li S, Zhou J, Xu T, Dou D and Xiong H 2022a Geomgcl: geometric graph contrastive learning for molecular property prediction *Proc. AAAI Conf. Artif. Intell.* **36** 4541–9

Li Z, Jiang M, Wang S and Zhang S 2022b Deep learning methods for molecular representation and property prediction *Drug Discov. Today* **27** 103373

Lin T, Wang Y, Liu X and Qiu X 2022 A survey of transformers *AI Open* **3** 111–32

Liu S, Nie W, Wang C, Lu J, Qiao Z, Liu L, Tang J, Xiao C and Anandkumar A 2022 Multi-modal molecule structure-text model for text-based retrieval and editing (arXiv:2212.10789)

Liu S, Wang H, Liu W, Lasenby J, Guo H and Tang J 2021 Pre-training molecular graph representation with 3d geometry (arXiv:2110.07728)

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D and Stoyanov V 2019 RoBERTa: a robustly optimized BERT pretraining approach (arXiv:1907.11692)

Lu C, Liu Q, Wang C, Huang Z, Lin P and He L 2019 Molecular property prediction: a multilevel quantum interactions modeling perspective *Proc. AAAI Conf. Artif. Intell.* **33** 1052–60

Martins I F, Teixeira A L, Pinheiro L and Falcao A O 2012 A Bayesian approach to in silico blood-brain barrier penetration modeling *J. Chem. Inf. Model.* **52** 1686–97

McInnes L, Healy J and Melville J 2018 Umap: uniform manifold approximation and projection for dimension reduction (arXiv:1802.03426)

Medical Dictionary for Regulatory Activities, MedDRA n.d. (available at: www.meddra.org/) (Accessed 20 February 2023)

Mobley D L and Guthrie J P 2014 FreeSolv: a database of experimental and calculated hydration free energies, with input files *J. Comput.-Aided Mol. Des.* **28** 711–20

Morris C, Ritzert M, Fey M, Hamilton W L, Lenssen J E, Rattan G and Grohe M 2019 Weisfeiler and Leman go neural: higher-order graph neural networks *Proc. AAAI Conf. Artif. Intell.* **33** 4602–9

Nigam A, Pollice R, Krenn M, Dos Passos Gomes G and Aspuru-Guzik A 2021 Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES *Chem. Sci.* **12** 7079–90

Oord A V D, Li Y and Vinyals O 2018 Representation learning with contrastive predictive coding (arXiv:1807.03748)

Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I 2019 *OpenAI Blog* (available at: https://life-extension.github.io/2020/05/27/GPT%E6%8A%80%E6%9C%AF%E5%88%9D%E6%8E%A2/language-models.pdf) (Accessed 9 June 2023)

Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y and Das P 2022 Large-scale chemical language representations capture molecular structure and properties *Nat. Mach. Intell.* **4** 1256–64

Russell R P 1988 Side effects of calcium channel blockers *Hypertension* **11** II42

Schaefer C P, Tome M E and Davis T P 2017 The opioid epidemic: a central role for the blood brain barrier in opioid analgesia and abuse *Fluids Barriers CNS* **14** 32

Schütt K, Kindermans P J, Sauceda Felix H E, Chmiela S, Tkatchenko A and Müller K R 2017 Schnet: a continuous-filter convolutional neural network for modeling quantum interactions *Advances in Neural Information Processing Systems* (*Long Beach, California, USA, 4–9 December 2017*) vol 30 (available at: https://proceedings.neurips.cc/paper/2017/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html)

Su J, Lu Y, Pan S, Murtadha A, Wen B and Liu Y 2021 Roformer: enhanced transformer with rotary position embedding (arXiv:2104.09864)

Subramanian G, Ramsundar B, Pande V and Denny R A 2016 Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches *J. Chem. Inf. Model.* **56** 1936–49

Tay Y, Dehghani M, Bahri D and Metzler D 2022 Efficient transformers: a survey *ACM Comput. Surv.* **55** 1–28

U.S. Department of Health and Human Services n.d. AIDS antiviral screen data—NCI DTP data—NCI Wiki (National Institutes of Health) (available at: https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data) (Accessed 20 February 2023)

Unsal S, Atas H, Albayrak M, Turhan K, Acar A C and Doğan T 2022 Learning functional properties of proteins with language models *Nat. Mach. Intell.* **4** 227–45

Vamathevan J *et al* 2019 Applications of machine learning in drug discovery and development *Nat. Rev. Drug Discovery* **18** 463–77

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* (*Long Beach, California, USA, 4–9 December 2017*) vol 30 (available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html)

Wang J, Krudy G, Hou T, Zhang W, Holland G and Xu X 2007 Development of reliable aqueous solubility models and their application in druglike analysis *J. Chem. Inf. Model.* **47** 1395–404

Wang N, Wang Y and Er M J 2022a Review on deep learning techniques for marine object recognition: architectures and algorithms *Control Eng. Pract.* **118** 104458

Wang R, Fang X, Lu Y and Wang S 2004 The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures *J. Med. Chem.* **47** 2977–80

Wang S, Guo Y, Wang Y, Sun H and Huang J 2019 SMILES-BERT: large scale unsupervised pre-training for molecular property prediction *Proc. 10th ACM Int. Conf. on Bioinformatics, Computational Biology and Health Informatics* pp 429–36

Wang Y, Wang J, Cao Z and Barati Farimani A 2022b Molecular contrastive learning of representations via graph neural networks *Nat. Mach. Intell.* **4** 279–87

Wen J *et al* 2023 Multimodal representation learning for predicting molecule–disease relations *Bioinformatics* **39** btad085

Wigh D S, Goodman J M and Lapkin A A 2022 A review of molecular representation in the age of machine learning *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **12** e1603

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A and Rush A M 2019 Huggingface's transformers: state-of-the-art natural language processing (arXiv:1910.03771)

Wu Z, Ramsundar B, Feinberg E N, Gomes J, Geniesse C, Pappu A S, Leswing K and Pande V 2018 MoleculeNet: a benchmark for molecular machine learning *Chem. Sci.* **9** 513–30

Xu K, Hu W, Leskovec J and Jegelka S 2018 How powerful are graph neural networks? (arXiv:1810.00826)

Xue D, Zhang H, Xiao D, Gong Y, Chuai G, Sun Y and Liu Q 2020 X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis *bioRxiv* (https://doi.org/10.1101/2020.12.23.424259) (Accessed 9 June 2023)

Yang K *et al* 2019 Analyzing learned molecular representations for property prediction *J. Chem. Inf. Model.* **59** 3370–88

Zhang S, Jin Y, Liu T, Wang Q, Zhang Z, Zhao S and Shan B 2022 SS-GNN: a simple-structured graph neural network for affinity prediction (arXiv:2206.07015)

Zheng L, Fan J and Mu Y 2019 Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction *ACS Omega* **4** 15956–65