Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Molformer: Large Scale Chemical Language Representations Capture Molecular Structure and Properties

Jerret Ross
IBM Research
Brian Belgodere
IBM Research
Vijil Chenthamarakshan
IBM Research https://orcid.org/0000-0001-7830-5777
Inkit Padhi
IBM Research
Youssef Mroueh
IBM https://orcid.org/0000-0001-8798-1267
Payel Das (daspa@us.ibm.com)
IBM Research https://orcid.org/0000-0002-7288-0516

Article

Keywords:

Posted Date: May 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1570270/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

MOLFORMER: Large-Scale Chemical Language Representations Capture Molecular Structure and Properties

Jerret Ross[1,*], **Brian Belgodere**[1], **Vijil Chenthamarakshan**[1], **Inkit Padhi**[1], **Youssef Mroueh**[1], **and Payel Das**[1,*]

[1]IBM Research, Yorktown Heights, NY 10598, USA
[*]rossja@us.ibm.com and daspa@us.ibm.com

ABSTRACT

Predicting the properties of a chemical molecule is of great importance in many applications, including drug discovery and material design. Machine learning-based models promise to enable more accurate and faster molecular property predictions than the current state-of-the-art techniques, such as Density Functional Theory calculations or wet-lab experiments. Various supervised machine learning models, including graph neural nets, have demonstrated promising performance in molecular property prediction tasks. However, the vast chemical space and the limited availability of property labels make supervised learning challenging, calling for learning a general-purpose molecular representation. Recently, unsupervised transformer-based language models pre-trained on large unlabeled corpus have produced state-of-the-art results in many downstream natural language processing tasks. Inspired by this development, we present molecular embeddings obtained by training an efficient transformer encoder model, MoLFormer, which uses rotary positional embeddings. This model employs a linear attention mechanism, coupled with highly distributed training, on SMILES sequences of 1.1 billion unlabeled molecules from the PubChem and ZINC datasets.Experiments show that utilizing the learned molecular representation outperforms existing baselines on downstream tasks, including supervised and self-supervised graph neural net baselines and language models, on several classification and regression tasks from ten benchmark datasets while performing competitively on two others. Further analyses, specifically through the lens of attention, demonstrate that MoLFormer trained on chemical SMILES indeed learns the spatial relationships between atoms within a molecule. These results provide encouraging evidence that the large-scale molecular language models can capture sufficient chemical and structural information to predict various distinct molecular properties, including quantum-chemical properties.

Main

Machine Learning (ML) has emerged as an appealing, computationally efficient approach for predicting molecular properties, with implications in drug discovery and material engineering. ML models for molecules can be trained directly on pre-defined chemical descriptors, such as unsupervised molecular fingerprints[1], or hand-derived derivatives of geometric features such as a Coulomb Matrix (CM)[2]. However, more recent ML models have focused on automatically learning the features either from the natural graphs that encode the connectivity information or from the line annotations of molecular structures, such as the popular SMILES[3] (Simplified Molecular-Input Line Entry System) representation. SMILES defines a character string representation of a molecule by performing a depth-first pre-order spanning tree traversal of the molecular graph, generating symbols for each atom, bond, tree-traversal decision, and broken cycles. Therefore, the resulting character string corresponds to a flattening of a spanning tree of the molecular graph. Learning on SMILES has been widely adopted for molecular property prediction[4–7] as SMILES is generally more compact than other methods of representing structure, including graphs. Additionally, meaningful substructures such as branches, cyclic structures, and chirality information are explicitly represented in SMILES strings, which is not the case for the graph representation.

However, the SMILES grammar is complex and restrictive; most sequences over the appropriate character set do not belong to well-defined molecules. Alternative string-based representations exist, such as SMARTS[8] and SELFIES[9]. Nevertheless, string-based representations are thought to not be topologically-aware, while graphs are. Due to these limitations, deep chemical language models may focus on learning the grammar of molecular strings and not the implicit topological structure of the molecular graphs. Accordingly, while string-based deep neural nets have been employed in predicting molecular properties[5–7,10], they are typically outperformed by graph neural networks (GNNs)[11] and their variants[12–20]. GNN frameworks can be generally viewed as "message passing", which includes local neighborhood information aggregation and information updates across different levels of granularity, e.g., nodes, edges, or the full graph, according to the graph's connectivity structure.

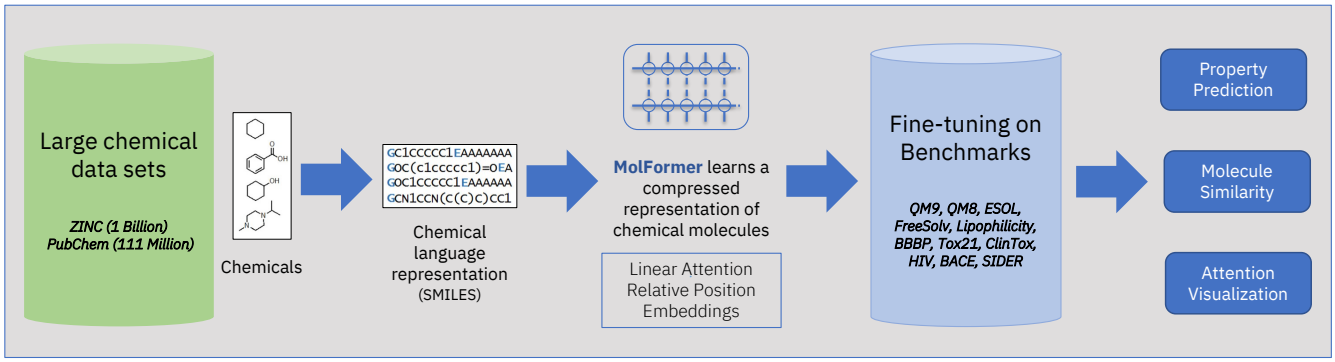

Figure 1. Overview of MoLFormer pipeline. The transformer neural network based model is trained on the SMILES sequences corresponding to a large collection of chemical molecules from PubChem and Zinc, two public chemical databases in a self-supervised fashion. MOLFORMER was designed with an efficient linear attention mechanism and relative positional embeddings, with the goal of learning a meaningful and compressed representation of chemical molecules. This foundation model was then adopted to different downstream molecular property prediction tasks via fine-tuning on task-specific data. The representative power was further tested by recovering molecular similarity using the MOLFORMER encodings, as well as by analyzing the correspondence between the interatomic spatial distance and attention value for a given molecule.

One challenge with supervised training of GNNs and language models for molecular property prediction is the scarcity of labeled data. Label annotation of molecules is typically expensive and this problem is compounded by the fact that the size of the space consisting of plausible chemicals in need of annotation is astronomically large (10^{60} to 10^{100})[21]. Such a scenario creates the need for molecular representation learning which can be generalizable to various property prediction tasks in an un-/self-supervised setting. The recent success of large transformer-based[22] foundation models[23], using the paradigm of learning a task-agnostic language representation, obtained by pre-training on large unlabeled corpora and subsequently using it for fine-tuning on downstream tasks of interest, has been extended to other domains.

Pre-trained Language Models (LMs)[24] and GNNs[25] have only recently started to emerge for predicting molecular properties. However, to what extent pre-trained LMs, trained on a large corpus of billions of molecules, are able to capture the molecule-property relationships across various downstream tasks remains unexplored.

Towards this direction, here we present molecular SMILES transformer models referred to as MOLFORMER (Molecular Language transFormer). We name our best performing MOLFORMER variant MOLFORMER-XL. MOLFORMER-XL was obtained using an efficient linear attention mechanism trained on a large corpus of 1.1 billion molecules (see Figure 1). Results show, for the first time, that pre-trained transformer encoders of molecular SMILES perform competitively with existing supervised or unsupervised LM and GNN baselines on predicting a wide variety of molecular properties, including quantum-mechanical properties.

Our main contributions are:

- We train a large-scale and efficient Molecular Language model transFormer (MOLFORMER) on over a billion molecules, with relatively limited hardware resources (up to 16 V100 GPUs). We owe our scalability and speedups to efficient linear time attention, adaptive bucketing of batches, and open-source parallelization provided in PyTorch Lightning and NCCL. With the combination of bucketing and linear attention we are able to achieve a batch size of 1600 molecules per GPU. Using 16 GPUs we need 208 hours to complete 4 epochs of pre-training for MOLFORMER-XL. To complete training in the same amount of time without bucketing and linear attention we would be limited to less than 50 molecules per GPU and require over 1000 GPUs for the task.
- We explore the difference between absolute and relative position embeddings in representing molecular SMILES. We also provide a new, efficient, and accurate linear attention approximation of the recently proposed relative position RoFormer[26].
- We perform extensive experimentation and ablation studies on several classification and regression tasks from 10 benchmark datasets, covering quantum mechanical, physical, biophysical, and physiological property prediction of small molecule chemicals from MoleculeNet[27].
- Our results provide encouraging evidence that MOLFORMER representations can accurately capture sufficient chemical and structural information to predict a diverse range of chemical properties. Furthermore, the performance of MOLFORMER is either better or on par with state-of-the-art GNNs that learn from precise graph topology information and beyond (e.g., bond distances).

- We provide further analyses to demonstrate that MOLFORMER can capture substructures, as well as spatial interatomic distances within a molecule from SMILES annotations only.

Related Work

Large Scale Training of Language Model The recent advancement of transformer-based masked language models (MLMs)[28,29] and prefix language models (PLMs)[30] have shown remarkable performance on various natural language understanding tasks. Self-supervised pre-trained representation learning of sequences through MLMs randomly masks input tokens during training and predicts these masked tokens, whereas PLMs require adding task-specific text tags to the input sequences. These language models show substantial performance improvements on downstream tasks via increasing transformer models size and pre-training using large-scale data corpora. Recent efforts have addressed the resulting cost and memory challenges encountered due to scaling up models and data. One such effort is the linear-time attention transformers introduced in[31–34] which address the quadratic memory challenges within the attention mechanism and allowing for more efficiency in training MLMs.

Molecular Representation Learning To represent molecules in vector space, traditional chemical fingerprints such as ECFP[1], have been used. Deep neural nets were further trained on chemical fingerprints for supervised learning. Recurrent Neural Network (RNN) based models have been used for molecular representation learning using SMILES and other linear molecular annotations as inputs[35]. At the same time, graph convolutional networks have been used to learn the neural fingerprints of molecules[11,36]. Previous work[17] implemented a single common framework to learn from graphs, referred to as a message passing framework, which computes node embeddings by aggregating neighborhood information during the message passing phase and computes a feature vector of the graph during the readout phase. Many attempts to extend GNNs have been made, which include variations of the original message passing concept to learn non-local effects; for instance, in[37] an attention mechanism was introduced. One challenge faced by GNNs is achieving higher expressivity that can distinguish between two given graphs to that of the hierarchy of the Weisfeiler-Lehman (WL) graph isomorphism, while maintaining scalability. It has been shown that typical message passing models have limited expressiveness and are not better than the first WL test (1-WL)[38]. Powerful deep models that represent higher order interactions between graph nodes have been suggested[38,39], but with a large increase in computational cost.

Molecular graphs can be further augmented with the 3D coordinates of atoms. Such augmentation is considered as privileged information due to the cost associated with deriving the 3D molecular geometry. To better model the spatial interactions among atoms the message passing framework was extended in[17] to include pairwise interatomic distances as edge features when geometric information was available. More recently, variations of the message passing networks (MPNN) were proposed to better model the spatial interactions within molecules and increase the models expressive power, e.g., by using continuous filter convolutional layers[40] or by using directional message passing[41] but at the cost of increased computational complexity. However, those models are not generalizable to settings where 3D structural information is not readily available and/or is expensive to compute (e.g. for larger molecules). Since the goal of this work is to learn a generalizable molecular representation from a large amount of unlabeled data without relying on expensive 3D information, we mainly focus on comparing the proposed MOLFORMER with existing supervised and un/self-supervised baselines that utilize different input representation (SMILES, graphs, fingerprints) and can be generalizable to a wide variety of tasks, from quantum mechanical to physiological.

Pre-trained Molecular Language and Graph Models The recent success of language representation models in downstream NLP tasks has inspired extending this paradigm to other domains. By combining the power of pre-training on large unlabeled corpus and contextual language models (LMs) using advanced neural nets, such as transformers, a domain-specific "language" embedding is obtained as the exclusive input for several downstream tasks.

Examples include understanding the language of life through advanced LMs trained on protein sequences. Here features extracted by LMs directly from single protein sequences reach state-of-the-art performance in downstream prediction tasks, even when those were used without evolutionary information[42–44]. Similar large-scale unsupervised pre-training on SMILES sequences have been explored for molecular property prediction[24,45–47]; however, those models did not attempt to predict a diverse range of molecular properties while exploiting the available chemical sequences at scale. Unsupervised/self-supervised representation learning has been tested on molecular graphs as well[25,48,49]. A more recent line of work has leveraged the power of contrastive self-supervised pre-training using 2D graph topology and 3D conformational geometry [50], which showed performance improvement on molecular regression and classification tasks compared to prior pre-training baselines. To our knowledge, the present study is the first one that explores the representational power of pre-trained chemical language models on predicting a broad range of downstream molecular properties from quantum chemical to physiological. In particular, predicting quantum-chemical properties from SMILES strings alone is non-trivial, as those properties are largely dependent on the 3D molecular geometry, which is considered privileged information and not available in general.

MoLFormer Framework

The goal of MOLFORMER is to learn a universal molecular representation from large scale chemical SMILES data and then evaluate the representation on various downstream molecular property prediction tasks, as shown in Figure 1. To do so, MOLFORMER model is developed using the masked language model framework, which randomly masks a certain percentage of tokens within a SMILES sequence during training and then predicts those tokens. The masked language modeling thus exploits self-supervision and enables contextual learning. Below we discuss details of the MOLFORMER model.

Model Details

As we aim to train a large scale masked language model of chemical SMILES efficiently and effectively, while utilizing relatively limited hardware resources, we leveraged transformer-based neural nets[22]. Transformers process inputs through a series of blocks alternating between self-attention and feed-forward connections. Transformers encode the position in the sequence via a positional embedding, termed the absolute positional embedding. The input feature at a position m is therefore concatenated with its corresponding absolute position embedding. Self-attention enables the network to construct complex representations that incorporate context from across the sequence. Attention mechanisms transform the features in the sequence into queries (q), keys (k), and value (v) representations. These representations produce the output of the attention at position m as follows:

$$\text{Attention}_m(Q,K,V) = \frac{\sum_{n=1}^{N} \exp(\langle q_m, k_n \rangle) v_n}{\sum_{n=1}^{N} \exp(\langle q_m, k_n \rangle)}.$$

A well known computational bottlenecks of the vanilla transformer[22] architecture is that the attention mechanism suffers from a quadratic computational cost with respect to the sequence length. Linear complexity attention models[31,51] have tackled this issue utilizing kernel approximations and random feature approximations variants. This led us to design MOLFORMER that utilizes an encoder based on a transformer with linear attention[51]. MOLFORMER with linear attention consists of 12 layers, 12 attention heads per layer, and has a hidden state size of 768. A Generalized Feature map[51] for the linear attention was chosen (see SI Section A.1.1 for details).

As mentioned above, in a transformer architecture the dependency between tokens at different position of a (chemical) sequence is modeled under the supervision of position encoding. The seminal work of[22] investigated absolute position embeddings to encode the position of a token in the sequence. More recent work[52–54] showed that use of relative position embeddings between tokens results in improved performance. Rotary position embeddings were introduced in RoFormer[26] as a means to enhance the relative encoding via position dependent rotations R_m of the query and the keys at a position m. These rotations can be efficiently implemented as pointwise multiplications and do not result in a dramatic computational increase.

In order to leverage Rotary embeddings with linear transformers, the use of the following approximation was proposed in[26]:

$$\text{Attention}_m(Q,K,V) = \frac{\sum_{n=1}^{N} \langle R_m \varphi(q_m), R_n \varphi(k_n) \rangle v_n}{\sum_{n=1}^{N} \langle \varphi(q_m), \varphi(k_n) \rangle},$$

where Q,K,V are the query, key, and value respectively, and φ a random feature map.

After preliminary experimentation with this linear Roformer, we found it performed worse than its absolute position counterpart. We propose the following modification to Roformer that we found to train more gracefully (the training loss falls faster and lower) than the original Roformer, as well as observing better performance than the model using absolute embeddings:

$$\text{Attention}_m(Q,K,V) = \frac{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle v_n}{\sum_{n=1}^{N} \langle \varphi(R_m q_m), \varphi(R_n k_n) \rangle}.$$

When compared with[26] we rotate the original keys and queries instead of the transformed ones with the feature map φ. We saw increased stability and faster convergence in training loss behavior when pre-training using rotary embeddings in contrast to absolute embeddings as observed in Figure 2.

Datasets and Tokenization

We constructed several datasets for pre-training by combining the PubChem[55] and ZINC[56] datasets with varying proportion from each. The PubChem dataset consists of 111 million molecules, while the much larger ZINC dataset contains over 1 billion molecules. To construct a vocabulary, we utilize the tokenizer from[57]. All molecules from both PubChem and ZINC are converted to a canonical format utilizing RDKit[58] then tokenized. All unique characters are extracted from the resulting output gives us a vocabulary of 2357 characters plus 5 special characters, resulting in a total of 2362 vocabulary characters which

are used for all pre-trained models considered in this paper, irrespective of pre-training dataset size. The post tokenization sequence length of the molecules range from 1 to just over 2000 characters. We decide to restrict the sequence length range from 1 character to 202 characters, special characters inclusive, to reduce computation time. Since over 99.4 percent of all molecules from our dataset contain less than 202 characters we hypothesize that the removal of molecules with more than 202 characters would be of minimal negative impact on pre-training.

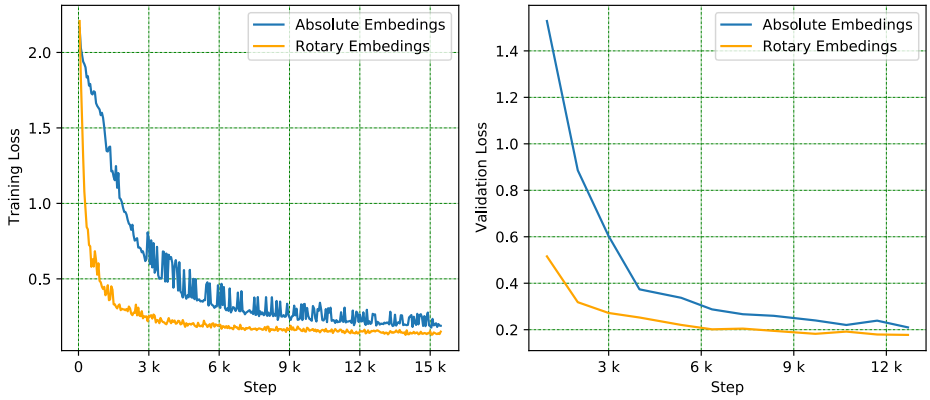

Figure 2. Training (Left) and validation (Right) losses of our Linear attention MOLFORMER with *rotary* (relative) and absolute position embeddings on PubChem. We see that both rotary and absolute MOLFORMER have graceful training curves. Our Rotary Linear attention MOLFORMER leads to lower training and validation losses than MOLFORMER with absolute position embeddings.

Large Scale Training and Parallelization

For pre-training we use the masked language model method defined in[29], 80% of the time a character will be replaced with the mask token, 10% of the time the character will be replaced with a random character and 10% of the time the character will be unchanged. Training was performed for 4 epochs through the entire PubChem+ZINC dataset with a fixed learning rate of $1.6e^{-4}$ and a batch size of 1600 molecules per GPU on a total of 16 GPUs over 2 servers connected via Infiniband fabric. It should be noted that as the number of GPUs utilized increased we found an increase in learning rate was necessary up to a factor of 8.

In order to scale our training to large datasets (1 Billion+ data points), we relied on adaptive bucketing of mini-batches by sequence length, as well as parallelization via distributed training (see Supplementary Information (SI) A for details). Using Linear attention and bucketing allowed us to reduce the number of GPUs needed from roughly 1000 for quadratic attention with no bucketing to 16.

Results and Discussion

To demonstrate the effectiveness of the pre-trained MOLFORMER as a universal and task-agnostic molecular representation, we benchmarked its adaptation performance on numerous challenging classification and regression tasks from MoleculeNet[27]. (Figure 1). Details of the benchmark datasets can be found in SI Section C.

Derivation of MOLFORMER Embeddings

We encode a chemical SMILES by extracting the mean of all embeddings of the last hidden state from the encoder model. The resulting embedding is used for all downstream tasks. The downstream tasks themselves can be divided into two categories, The first category being called *Frozen* and the second being called *Fine-tuned*. The *Frozen* setting is defined by training a fully connected model for each task, while keeping the encoder embeddings fixed. The second setting, *Fine-tuned*, involves fine-tuning the weights of the encoder model jointly with the fully connected model for each downstream task. The ideal configuration and hyperparameters for the frozen strategy are discovered through a grid search as described in SI Table 8. For the fine-tuned strategy, we use a 2-layer fully connected network with a hidden dimension of 768 (matching the encoder embedding) with Dropout (set to 0.1) and GELU layers in-between, on top of a final single output dimension for regression tasks.

Performance of MOLFORMER Embeddings on Downstream Tasks

We evaluate the performance of MOLFORMER embeddings and compare them with existing baselines on six classification and five regression tasks from the MoleculeNet benchmark[27], as discussed below. We refer to MOLFORMER which has been pre-trained on the entire training set comprised of ≈ 1.1 B molecules (all molecules from both PubChem and Zinc) as MOLFORMER-XL. Unless stated otherwise, the MOLFORMER-XL is trained with linear attention using rotary positional embeddings and the performance reported is of the model fine-tuned on the downstream task. To predict various properties on the downstream tasks we fined-tuned the model as described in the previous section. We use the training, validation and testing data split as defined by the MoleculeNet benchmark for all tasks (see SI C).

Dataset	BBBP	Tox21	ClinTox	HIV	BACE	SIDER
Tasks	1	12	2	1	1	27
RF	71.4	76.9	71.3	78.1	**86.7**	**68.4**
SVM	72.9	**81.8**	66.9	**79.2**	86.2	68.2
MGCN[59]	**85.0**	70.7	63.4	73.8	73.4	55.2
D-MPNN[60]	71.2	68.9	**90.5**	75.0	85.3	63.2
Hu, et al.[61]	70.8	78.7	78.9	80.2	85.9	65.2
N-Gram[48]	91.2	76.9	85.5	**83.0**	87.6	63.2
MolCLR[25]	73.6	79.8	93.2	80.6	**89.0**	68.0
GraphMVP-C[50]	72.4	74.4	77.5	77.0	81.2	63.9
ChemBerta[24]	64.3	-	90.6	62.2	-	-
MOLFORMER-XL	**93.7**	**84.7**	**94.8**	82.2	88.21	**69.0**

Table 1. Comparison of fine-tuned MOLFORMER with existing supervised and pre-trained/self-supervised baselines on multiple classification benchmarks. All models were evaluated by AUC-ROC on scaffold splits. Baseline performances are adopted from references[24,25,50].

Classification Tasks We choose six classification tasks from the MoleculeNet benchmark with nine total baselines, four supervised and five self-supervised, for comparison against MOLFORMER-XL. The supervised baselines consist of shallow machine learning models trained on molecular fingerprints (RF and SVM in Table 1) or graph neural nets. Among the pre-trained/self-supervised baselines, Hu, et al.[61] pre-trains a Graph Isomorphism Network (GIN, a GNN that uses an MLP and weighted sum of node features in the aggregation) on molecular graphs that includes edge features involved in aggregation. N-gram graph[48] uses a simple unsupervised representation for molecules by first embedding the nodes in a graph and then constructing a compact representation of the graph by assembling the vertex embeddings in short walks in the graph. MolCLR[25] is a self-supervised learning framework based on GIN, which uses contrastive loss[62,63]. GraphMVP-C is the Graph Multi-View Pre-training (GraphMVP) framework proposed by [50], where self-supervised learning (SSL) is performed by leveraging the correspondence and consistency between 2D topological structures and 3D geometric views. ChemBerta[24] is a pre-trained molecular language model trained on a smaller chemical dataset. Table 1 documents the performance comparison of MOLFORMER with these baselines on six classification benchmarks using the MoleculeNet scaffold data splits. MOLFORMER-XL outperforms all baselines in four (BBBP, Tox21, ClinTox, and SIDER) out of six benchmarks and comes a close second in the remaining two (HIV and BACE).

Regression Tasks Next, we evaluate MOLFORMER-XL on more challenging regression tasks from MoleculeNet. We report our performance on five regression benchmarks, namely QM9, QM8, ESOL, FreeSolv, and Lipophilicity, in Table 2. In particular, QM9 and QM8 involve predicting several quantum chemical measures, which is considered challenging without having access to privileged 3D geometric information. Again we use the train, validation and test split as suggested in[27] for these tasks. The baselines considered are a molecular graph convolutional network (GC, a GNN that utilizes a mean-pooling over the node and its neighbors before the linear transformation)[64], the attentive-FP (A-FP) model[37], and an MPNN variant[17] that learns edge features such as pairwise interatomic distances. Results show that MOLFORMER-XL upon task-specific fine-tuning outperforms the existing supervised GNN baselines, specifically GC, A-FP, and MPNN (augmented with bond distances for QM8 and QM9), on all five tasks. These results, combined with MOLFORMER-XL performance on the classification benchmarks confirm its generalizability.

A Closer Look at QM9 We further report MOLFORMER-XL performance on all twelve property prediction tasks individually within QM9 (see SI Table 9), and compare that against several previously discussed baseline models as well as four additional baselines. The additional baselines included are as follows: (i) a more expressive GNN, specifically 123-GNN[38], (ii) two neural nets that leverage 3D geometry – a multitask neural net encoding the Coulomb Matrix (CM)[65] and its GNN variant as in the deep

Task	GC	A-FP	MPNN	MoLFormer-XL
QM9 (avg MAE)	4.3536	2.6355	3.1898	**1.5894**
QM8 (avg MAE)	0.0148	0.0282	0.0143	**0.0102**
ESOL (RMSE)	0.97	0.503	0.58	**0.2787**
FreeSolv (RMSE)	1.40	0.736	1.15	**0.2308**
Lipophilicity (RMSE)	0.655	0.578	0.719	**0.5289**

Table 2. Performance of fine-tuned MoLFormer and other supervised GNN baselines on QM9, QM8, ESOL, FreeSolv, and Lipophilicity regression benchmarks. Baseline performances are taken from references[27,37].

tensor neural net (DTNN)[66], and (iii) Chemberta[24]. The results are reported in SI Table 14, which show that MoLFormer-XL achieves comparable or better performance to that of the majority of the competitors. Specifically, MoLFormer-XL outperforms all baselines in term of average MAE and average standard MAE. ChemBERTa shows the highest average MAE among all. The more expressive 123-GNN performs better on most measures compared to MoLFormer-XL; however, such powerful networks are known to be difficult to scale (see[67] for example). As a comparison, the linear attention employed in MoLFormer-XL ensures a linear time complexity.

The motivation for the increased focus on individual QM9 tasks in this section is to put into question the consensus in the community that the SMILES representation does not explicitly encode topological or geometric information, which is needed for accurately capturing quantum-mechanical properties. If this consensus is in fact correct then a model, such as MoLFormer-XL, will perform poorly on QM9 tasks when trained on only SMILES representation. As stated above MoLFormer-XL, with task-specific fine-tuning trained on SMILES strings from a large molecular corpus achieves comparable or better performance to that of the majority of the competitors. The performance MoLFormer-XL achieves on the QM9 tasks casts doubt on this consensus.

Ablation Studies

In this section we discuss several different ablations of MoLFormer-XL in an attempt to provide insights into its impressive performance. The ablations we performed can be broadly divided in the following three categories (1) the effect of size and the nature of the pre-training data, and (2) the results with (*frozen*) and with (*fine-tuned*) model fine-tuning on the downstream data, (3) the effect of absolute and rotary positional embeddings.

Data Size First we investigate how pre-training dataset size affects the performance of MoLFormer-XL on several downstream tasks from the MoleculeNet benchmark. To accomplish this we chose 3 different weighted combinations of the PubChem and Zinc datasets, specifically a set consisting of 10% of Zinc and 10% PubChem, another with 100% of PubChem mixed with 10% of Zinc, and then one with 100% Zinc molecules and 0% PubChem. All models are pre-trained with rotary embeddings and linear attention and then compared to MoLFormer-XL. Identical learning rates, data splits, optimization, etc. are used for pre-training and fine-tuning. Tables 3 and 4 summarize these results. While MoLFormer-XL performs better on average, we report two interesting observations. The first is that the model that is pre-trained on the second biggest data set, 100% Zinc, consistently performs worse than all other pre-trained models. The other point of interest is that when MoLFormer-XL falls behind, it is only by a very small margin (See performance on ESOL, QM8, FreeSolv benchmarks in Table 3).

Task	10% ZINC + 10% PubChem	100% PubChem + 10% ZINC	100% ZINC	MoLFormer-XL
QM9 (avg MAE)	1.7754	1.9093	1.9403	**1.5894**
QM8 (avg MAE)	0.0108	**0.0102**	0.0124	**0.0102**
ESOL (RMSE)	0.3295	**0.2775**	0.3023	0.2787
FreeSolv (RMSE)	0.2221	**0.2050**	0.2981	0.2308
Lipophilicity (RMSE)	0.5472	0.5331	0.5400	**0.5289**

Table 3. Performance comparison of fine-tuned MoLFormer-XL with fine-tuned MoLFormer models pre-trained on smaller datasets on QM9, QM8, ESOL, FreeSolv and Lipophilicity regression benchmarks.

Fine-tuned* versus *Frozen Table 5 further summarizes the two remaining ablation experiments using the QM9 benchmark. For simplicity we observe that the the *fine-tuned* ablation experiments achieves such a convincing win over the *frozen* experiments on all pre-training dataset sizes that we opted to only investigate fine-tuning for all other benchmarks.

Dataset	BBBP	HIV	BACE
10% ZINC + 10% PubChem	91.5	81.3	86.6
100% ZINC + 10% PubChem	92.2	79.2	86.3
100% ZINC	89.9	78.4	87.7
MoLFormer-XL	**93.7**	**82.2**	**88.2**

Table 4. Comparison of MoLFormer-XL with fine-tuned MoLFormer models pre-trained on smaller datasets on BBBP, HIV, and BACE classification benchmarks.

Pre-training Data →	QM9 Only			PubChem Only			PubChem+ZINC		
Dataset Size →	111×10^3			111×10^6			$> 1.1 \times 10^9$		
Measure ↓	Frozen × Rotary	Fine-tuned × Rotary	Fine-tuned ✓ Rotary	Frozen × Rotary	Fine-tuned × Rotary	Fine-tuned ✓ Rotary	Frozen × Rotary	Fine-tuned × Rotary	Fine-tuned ✓ Rotary
Avg MAE	8.3808	**2.4621**	2.6604	8.2600	**2.9680**	3.3990	2.5497	1.8620	**1.5894**
Avg std MAE	0.2390	**0.0843**	0.0937	0.2447	**0.0801**	0.1355	0.0978	0.0611	**0.0567**

Table 5. Comparison of different MoLFormer variants on QM9 test set, in terms of average MAE and average standard MAE. Variants considered are MoLFormer pre-trained using QM9 only, PubChem only, and PubChem+ZINC dataset. The variants with and without fine-tuning on downstream task are compared, as well as models with, (✓)Rotary, and without ,(×)Rotary, rotary embeddings. Our best candidate variant (for Table 14) is chosen based on the average MAE (Mean Absolute Error) score, lower is better.

Position embeddings The positional embeddings ablation results are collected in Table 5. Different pre-training datasets sizes are also investigated and are broken up into MoLFormer (1) pre-trained on only the QM9 training set (111k molecules) – referred to as MoLFormer-qm9 (2) only PubChem (111M molecules) – referred to as MoLFormer-PubChem (3) PubChem+ZINC (1.1 Billion+ Molecules), *i.e.* MoLFormer-XL. Results presented in Table 5 show that MoLFormer with Rotary embeddings and fine-tuning are behind the Absolute positional embedding model for the smaller datasets, but then wins as the dataset size passes 1 Billion molecules. The underlying reason for this observation will be investigated in future work. With that said we can see that as the pre-training grows from PubChem only to the more extensive and diverse PubChem+Zinc corpus the representational power of the model increases, as showcased by the stronger performance on the QM9 benchmark.

Robustness Across Data Folds We report performance comparison at the individual property prediction task within QM9 in SI Table 15. In order to ensure the robustness of these results across data splits we also provide, in the SI, the performance of MoLFormer-XL on QM9 tasks using 5-cross validation folds (SI Table 16).

Insights into MoLFormer

We performed the following set of experiments in order to evaluate if the MoLFormer embeddings capture molecular properties as well as structure. First, we investigate a t-SNE[68] projection of MoLFormer-XL embeddings sampled from two classes of the BBBP dataset as shown in Figure 3. These two classes contain molecules that penetrate (penetrating) the blood brain barrier and molecules that do not penetrate (non-penetrating) the blood brain barrier. It can be seen in Figure 3, that even without task-specific fine-tuning, MoLFormer-XL is able to discriminate between the two classes, suggesting that molecular property information has been captured in this universal representation. Next, we analyze the 2D graph and 3D geometric information content in MoLFormer-XL embeddings.

Molecular Similarity Recovery

Next, we analysed the correlation between pairwise similarities estimated using the Tanimoto distance, a popular measure of pairwise distance between chemicals, on the molecular fingerprints and those estimated using the Euclidean distance on the MoLFormer-XL embeddings. We further looked into the correlation between the number of atoms in the maximum common subgraph of a pair of molecules with their corresponding euclidean distance in the embedding space for a set of random molecules picked from PubChem. The results are summarized in Table 6 and show that MoLFormer-XL embeddings are better correlated with known molecule similarity measures when compared to ChemBERTa. These results are suggestive of MoLFormer embeddings being informative of chemical structure similarity.

Attention Analyses

Finally, we inspect the average pooled attention matrices of MoLFormer-XL to explore the chemical information embedded in them. For this purpose, we utilize the cosine similarities between attention values and the spatial distances between atoms

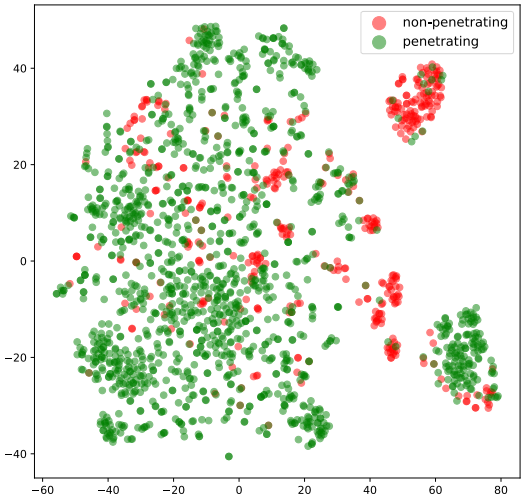

Figure 3. t-SNE projection of frozen MOLFORMER embeddings (no fine-tuning) of the BBBP dataset.

Correlation	ChemBERTa	MOLFORMER-XL
Fingerprint	0.48	**0.64**
MCS	-0.44	**-0.60**

Table 6. Correlation with structural similarity metrics on 10000 pairs of molecules from the PubChem dataset. Reported correlations are between (1) the pairwise similarities estimated using molecular Fingerprints and those using MOLFORMER-XL (or ChemBERTa) embeddings and (2) the number of atoms in the maximum common subgraph (MCS) of two molecules and their corresponding Euclidean distance in the embedding space.

Distance-Category	Attention	1	3	5	7	9	11
Short	Full (✓ Rotary)	**0.615**	0.604	0.603	**0.615**	0.601	**0.598**
	Linear (✓ Rotary)	0.596	0.597	0.602	0.597	0.600	0.594
Medium	Full (✓ Rotary)	0.716	0.724	0.724	0.716	**0.727**	0.727
	Linear (✓ Rotary)	**0.729**	**0.728**	**0.724**	**0.727**	0.726	**0.730**
Long	Full (✓ Rotary)	0.204	0.207	0.208	0.205	0.208	0.210
	Linear (✓ Rotary)	**0.211**	**0.210**	**0.210**	**0.211**	**0.209**	**0.210**

Table 7. Comparison of MOLFORMER models with respect to cosine similarity between the interatomic spatial distance map and the attention map, across three different distance categories for 7806 molecules from QM9 test set. Short, Medium, and Long distance categories are defined with interatomic distances in the range of $\leq$2, 2-4, and 4-10 Å, respectively.

within a molecule from the QM9 test set. Spatial distances are obtained from the corresponding energy-minimized geometries provided within QM9 benchmark[69]. MOLFORMER-XL is compared with a MOLFORMER variant trained with full attention and rotary embeddings on the entire PubChem+Zinc dataset. Note that the MOLFORMER models here are not fine-tuned for the QM9 dataset. The frozen MOLFORMER with full attention shows a much higher average MAE ($\geq$ 12) on QM9 downstream tasks, performance is particularly worse on internal energies (U and U_0), enthalpy (H), and free energy (G). We present attention results separately for three different categories of interatomic spatial distances: short ($\leq$ 2 Å; that are mostly reflective of typical covalent bonds in the molecule, C-C single bond distance being 1.5 Å), medium (2-4 Å) and long ($\geq$ 4Å), and summarize them in Table 7. Interestingly, attentions in MOLFORMER with linear or full attention (and rotary positional embeddings) show strong similarity with interatomic distances in both the short and medium categories, while revealing a weak (around 0.2) similarity with longer interatomic distances. This is an interesting observation, indicating that MOLFORMER is able to capture spatial relations between atomic tokens that are not necessarily neighbors in the SMILES sequence. The observed attentions in MOLFORMER-XL are slightly more in line with medium and long range distances, when compared to MOLFORMER with full attention. This observation suggests MOLFORMER-XL, with linear-attention, does in fact capture spatial relations between atoms more effectively.

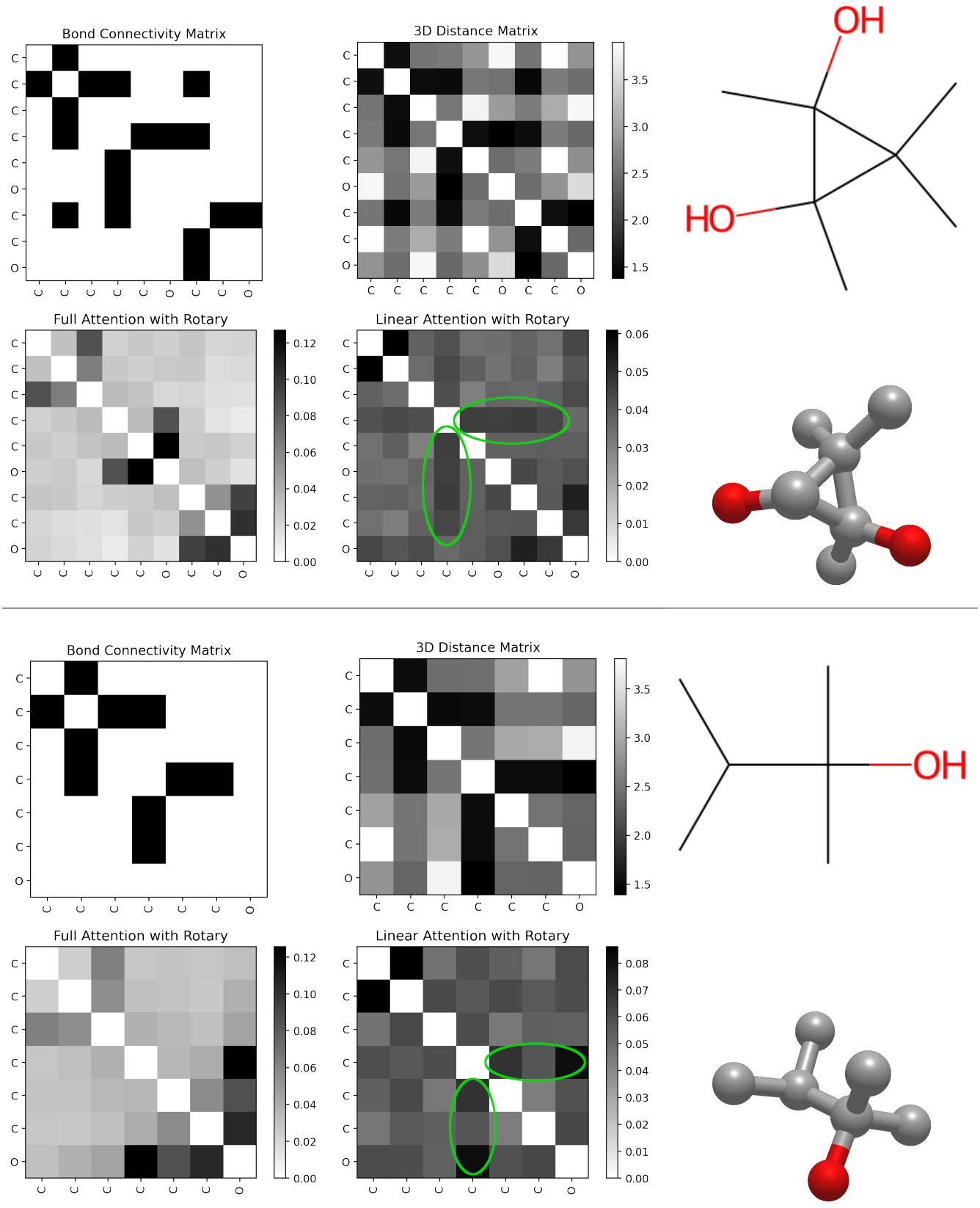

Figure 4. Visualization of the learned attention map (using either full or linear attention) under rotary embedding and corresponding molecular structure (bond connectivity and 3D distance) for two random molecules: `'CC1(C)C(C)(O)C1(C)O'` (top) and `'CC(C)C(C)(C)O'` (bottom). The attention map (only tokens that map to constituent atoms are shown for clarity), comprised of the average-pooled heads of an intermediate attention layer, exhibits awareness of both covalent bond connectivity and interatomic long-range spatial relationship. The linear attention variant captures (encircled in green) the medium 3D range distance better in comparison to its counterpart.

Figure 4 further elaborates this point showing the average learned attention coefficients in an intermediate attention layer of MOLFORMER-XL with rotary positional embeddings. Attentions between different pairs of atomic tokens are compared to the corresponding covalent bond connectivity and 3D distances between atom pairs (complete attention matrices for the same molecules across all layers are shown in Figures 8 and 9 in SI). We chose two molecules from the QM9 test set whose attention values show a high cosine similarity with the medium range spatial distances for this visualization. Visual inspection indicates that an aggregation of heads on the intermediate rotary attention layer corresponds well to the covalent bonding pattern, while also capturing the signature of the spatial relations between non-bonded atoms within a molecule. These attention analysis results suggest that MOLFORMER-XL is able to recover molecular structural information from corresponding SMILES sequence to a significant extent. This capability likely stems from pre-training on a large corpus of chemical SMILES which also allows MOLFORMER-XL to learn fundamental properties of chemicals, including structural information and various downstream properties, ranging from quantum chemical to physiological. A similar observation has been reported in recent work on protein sequence modeling[42,43]. To our knowledge, this is the first confirmation that structural and diverse property information emerges in the representation learned by a chemical language model pre-trained on large-scale data.

Conclusion

In this work, we have explored the power of unsupervised large-scale pre-trained molecular language models at various molecular property prediction tasks. Unlike graphs, molecular languages such as SMILES do not explicitly encode molecular topology. However, with well-designed self-supervised training on a large-scale corpus and with an expressive architecture, such as a contextualized transformer-based language model with a linear attention mechanism, and a parallelized training protocol, our MOLFORMER can efficiently learn implicit rich structure-property relationship information.

Specifically, MOLFORMER outperforms existing graph-based baselines on a wide variety of molecular regression and classification benchmarks. To our knowledge, this is the first work that validates the power of large-scale self-supervised pre-trained molecular language models on predicting molecular properties across the entire range from quantum chemical to physiological. Further, by analysing the learned attentions, we show that MOLFORMER trained on SMILES sequences indeed is aware of interatomic relations within a molecule, even beyond the 2D topology. Finally, on the large-scale learning end, we showcased with MOLFORMER an efficient and environment-friendly use of computational resources, reducing the number of GPUs needed to perform the training by a factor of 60 (1000 vs. 16).

MOLFORMER has immediate potential for faster in silico screening of molecules across diverse targets, which is important for material design and drug discovery applications with positive societal impact. However, it should be noted that misuse of such technology without a proper experimental and scientific validation in a wet lab can have harmful implications. Further, it has been shown that accurate property prediction models (for example., for predicting toxicity) along with generative models can be exploited for designing highly toxic molecules[70]. This highlights the need for a responsible framework around the use of these emerging powerful technologies. In addition, the present work calls for further exploration of the representational power of MOLFORMER in the context of its ability to learn structural molecular information directly from chemical language and can be extended beyond the small organic molecules studied in this work. Future work will also aim to improve MOLFORMER by employing larger models and larger training data, using improved and/or domain-specific self-supervised tasks, and using other string-based representations like SELFIES[9].

References

1. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. chemical information modeling* **50**, 742–754 (2010).
2. Rupp, M., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. review letters* **108**, 058301 (2012).
3. Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. chemical information computer sciences* **28**, 31–36 (1988).
4. Goh, G. B., Hodas, N. O., Siegel, C. & Vishnu, A. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv:1712.02034* (2017).
5. Öztürk, H., Özgür, A. & Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).
6. Paul, A. *et al.* Chemixnet: Mixed dnn architectures for predicting chemical properties using multiple molecular representations. *arXiv:1811.08283* (2018).
7. Shin, B., Park, S., Kang, K. & Ho, J. C. Self-attention based molecule representation for predicting drug-target interaction. In *Machine Learning for Healthcare Conference*, 230–248 (PMLR, 2019).

8. Daylight Chemical Information Systems, I. Smarts™—a language for describing molecular patterns (2007).
9. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **1**, 045024, DOI: 10.1088/2632-2153/aba947 (2020).
10. Jo, J., Kwak, B., Choi, H.-S. & Yoon, S. The message passing neural networks for chemical property prediction on smiles. *Methods* **179**, 65–72 (2020). Interpretable machine learning in bioinformatics.
11. Duvenaud, D. *et al.* Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15 (2015).
12. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering (2017). 1606.09375.
13. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks (2017). 1609.02907.
14. Li, Y., Tarlow, D., Brockschmidt, M. & Zemel, R. Gated graph sequence neural networks. *arXiv:1511.05493* (2015).
15. Veličković, P. *et al.* Graph attention networks. *arxiv, 1710.10903* (2018).
16. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs (2018). 1706.02216.
17. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry (2017). 1704.01212.
18. Schlichtkrull, M. *et al.* Modeling relational data with graph convolutional networks (2017). 1703.06103.
19. Liao, R., Zhao, Z., Urtasun, R. & Zemel, R. S. Lanczosnet: Multi-scale deep graph convolutional networks (2019). 1901.01484.
20. Chen, P., Liu, W., Hsieh, C.-Y., Chen, G. & Zhang, S. Utilizing edge features in graph neural networks via variational information maximization. *1906.05488* (2019).
21. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823–824 (2004).
22. Vaswani, A. *et al.* Attention is all you need. *arXiv:1706.03762* (2017).
23. Bommasani, R. *et al.* On the opportunities and risks of foundation models, DOI: 10.48550/ARXIV.2108.07258 (2021).
24. Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction (2020). 2010.09885.
25. Wang, Y., Wang, J., Cao, Z. & Farimani, A. B. Molclr: Molecular contrastive learning of representations via graph neural networks (2021). 2102.10056.
26. Su, J., Lu, Y., Pan, S., Wen, B. & Liu, Y. Roformer: Enhanced transformer with rotary position embedding (2021). 2104.09864.
27. Wu, Z. *et al.* Moleculenet: A benchmark for molecular machine learning (2018). 1703.00564.
28. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692* (2019).
29. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the NAACL: HLT, Vol 1* (2019).
30. Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR* (2020).
31. Choromanski, K. *et al.* Rethinking attention with performers. *CoRR* **abs/2009.14794** (2020).
32. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The long-document transformer. *arXiv:2004.05150* (2020).
33. Kitaev, N., Kaiser, L. & Levskaya, A. Reformer: The efficient transformer. In *ICLR* (2020).
34. Wang, S., Li, B. Z., Khabsa, M., Fang, H. & Ma, H. Linformer: Self-attention with linear complexity (2020). 2006.04768.
35. Bjerrum, E. J. Smiles enumeration as data augmentation for neural network modeling of molecules (2017). 1703.07076.
36. Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. chemical information modeling* **57**, 1757–1772 (2017).
37. Xiong, Z. *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. medicinal chemistry* (2019).
38. Morris, C. *et al.* Weisfeiler and leman go neural: Higher-order graph neural networks. In *AAAI*, vol. 33, 4602–4609 (2019).
39. Maron, H., Ben-Hamu, H., Serviansky, H. & Lipman, Y. Provably powerful graph networks (2020). 1905.11136.

40. Schütt, K. T. *et al.* Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *NeurIPS* (2017).

41. Klicpera, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs (2020). 2003.03123.

42. Vig, J. *et al.* Bertology meets biology: Interpreting attention in protein language models. *arXiv:2006.15222* (2020).

43. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv* DOI: 10.1101/622803 (2020).

44. Elnaggar, A. *et al.* Prottrans: Towards cracking the language of life's code through self-supervised learning. *bioRxiv* DOI: 10.1101/2020.07.12.199554 (2021).

45. Xue, D. *et al.* X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis. *bioRxiv* DOI: 10.1101/2020.12.23.424259 (2020).

46. Wang, S., Guo, Y., Wang, Y., Sun, H. & Huang, J. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM-BCB* (2019).

47. Kim, H., Lee, J., Ahn, S. & Lee, J. R. A merged molecular representation learning for molecular properties prediction with a web-based service. *Sci. Reports* **11**, 1–9 (2021).

48. Liu, S., Demirel, M. F. & Liang, Y. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules (2019). 1806.09206.

49. Rong, Y. *et al.* Self-supervised graph transformer on large-scale molecular data (2020). 2007.02835.

50. Liu, S. *et al.* Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728* (2021).

51. Katharopoulos, A., Vyas, A., Pappas, N. & Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linearattention. *CoRR* **abs/2006.16236** (2020). 2006.16236.

52. Shaw, P., Uszkoreit, J. & Vaswani, A. Self-attention with relative position representations. In *NAACL-HLT*, 464–468 (Association for Computational Linguistics, New Orleans, Louisiana, 2018).

53. Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR* **21**, 1–67 (2020).

54. Ke, G., He, D. & Liu, T.-Y. Rethinking positional encoding in language pre-training. In *ICLR* (2021).

55. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* (2018).

56. Irwin, J. J. & Shoichet, B. K. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).

57. Schwaller, P. *et al.* Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583, DOI: 10.1021/acscentsci.9b00576 (2019).

58. RDKit: Open-source cheminformatics. http://www.rdkit.org (2021). [Online; accessed 28-May-2021].

59. Lu, C. *et al.* Molecular property prediction: A multilevel quantum interactions modeling perspective. In *AAAI*, 1052–1060, DOI: 10.1609/aaai.v33i01.33011052 (AAAI Press, 2019).

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

61. Hu, W. *et al.* Strategies for pre-training graph neural networks (2020). 1905.12265.

62. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations (2020). 2002.05709.

63. van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding (2019). 1807.03748.

64. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS central science* (2017).

65. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301, DOI: 10.1103/PhysRevLett.108.058301 (2012).

66. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. communications* **8**, 1–8 (2017).

67. Vignac, C., Loukas, A. & Frossard, P. Building powerful and equivariant graph neural networks with structural message-passing (2020). 2006.15107.

68. van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

69. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).

70. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nat. Mach. Intell.* **4**, 189–191, DOI: 10.1038/s42256-022-00465-9 (2022).

71. You, Y. *et al.* Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv:1904.00962* (2019).

72. Falcon, e. a., WA. Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning* **3** (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MolFormernaturesupplement.pdf
- MolFormerNatMatIntelSI.pdf