nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00580-7

Large-scale chemical language representations capture molecular structure and properties

Received: 18 April 2022

Accepted: 3 November 2022

Published online: 21 December 2022

Check for updates

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh & Payel Das

Models based on machine learning can enable accurate and fast molecular property predictions, which is of interest in drug discovery and material design. Various supervised machine learning models have demonstrated promising performance, but the vast chemical space and the limited availability of property labels make supervised learning challenging. Recently, unsupervised transformer-based language models pretrained on a large unlabelled 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 unlabelled molecules from the PubChem and ZINC datasets. We show that the learned molecular representation outperforms existing baselines, including supervised and self-supervised graph neural networks and language models, on several downstream tasks from ten benchmark datasets. They perform 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 large-scale molecular language models can capture sufficient chemical and structural information to predict various distinct molecular properties, including quantum-chemical properties.

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 predefined chemical descriptors, such as unsupervised molecular fingerprints[1], or hand-derived derivatives of geometric features such as a Coulomb matrix[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 preorder spanning tree traversal of the molecular graph, generating symbols for each atom, bond, tree-traversal decision and broken cycle. 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

IBM Research, Yorktown Heights, New York, NY, USA. e-mail: rossja@us.ibm.com; daspa@us.ibm.com

https://doi.org/10.1038/s42256-022-00580-7

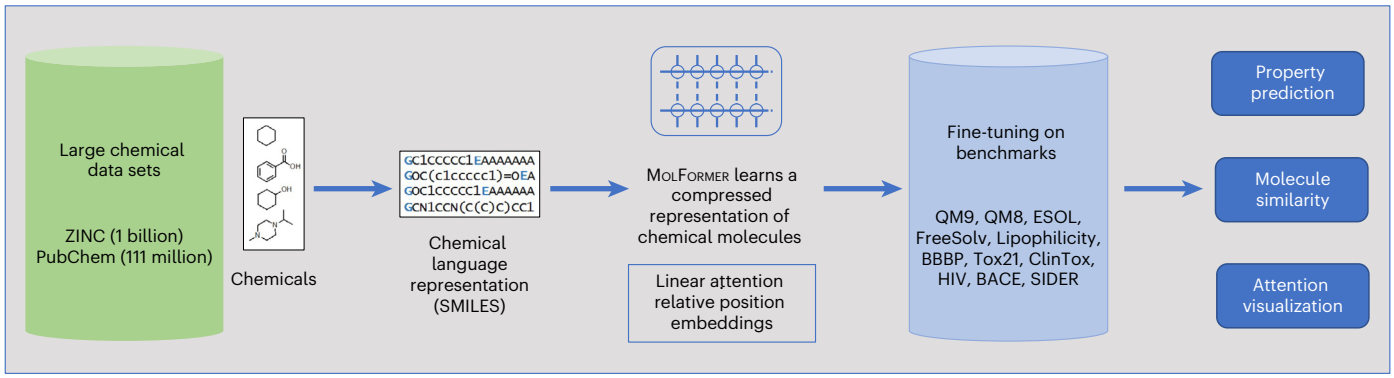

Fig. 1 | Overview of MolFormer pipeline. The transformer neural network based model is trained on the SMILES sequences corresponding to a large collection of chemical molecules from PubChem and ZINC, two public chemical databases, in a self-supervised fashion. MolFormer was designed with an efficient 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 adapted 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 analysing the correspondence between the interatomic spatial distance and attention value for a given molecule.

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]. Comparing benefits of these alternative representations with respect to SMILES is an active area of research. For example, ref. 10, focusing on molecular optimization tasks on the learned representation space, suggested no obvious shortcoming of SMILES with respect to SELFIES in terms of optimization ability and sample efficiency, particularly when the language model is more advanced. 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[5–7,11], they are typically outperformed by graph neural networks (GNNs)[12] and their variants[13–21]. GNN frameworks can be generally viewed as 'message passing', which includes local neighbourhood information aggregation and information updates across different levels of granularity, for example, nodes, edges or the full graph, according to the graph's connectivity structure.

One challenge with supervised training of GNNs and language models for molecular property prediction is the scarcity of labelled 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})[22]. Such a scenario creates the need for molecular representation learning that can be generalizable to various property prediction tasks in an un-/self-supervised setting. The recent success of large transformer-based[23] foundation models[24], using the paradigm of learning a task-agnostic language representation, obtained by pretraining on large unlabelled corpora and subsequently using it for fine-tuning on downstream tasks of interest, has been extended to other domains.

Pretrained language models[25] and GNNs[26] for predicting molecular properties have only recently started to emerge. However, to what extent pretrained language models, 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 (Fig. 1). Results show that pretrained transformer encoders of molecular SMILES perform competitively with existing supervised or unsupervised language model and GNN baselines in predicting a wide variety of molecular properties, including quantum-mechanical properties.

Our main contributions are the following.

- 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 graphics processing units (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 1,600 molecules per GPU. Using 16 GPUs we need 208 h to complete four epochs of pretraining for MolFormer-XL. To complete training in the same amount of time without bucketing and linear attention we would be limited to fewer than 50 molecules per GPU and require over 1,000 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[27].
- We perform extensive experimentation and ablation studies on several classification and regression tasks from ten benchmark datasets, covering quantum-mechanical, physical, biophysical and physiological property prediction of small-molecule chemicals from MoleculeNet[28].
- 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 than or on a par with state-of-the-art GNNs that learn from precise graph topology information and beyond (for example, 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.

https://doi.org/10.1038/s42256-022-00580-7

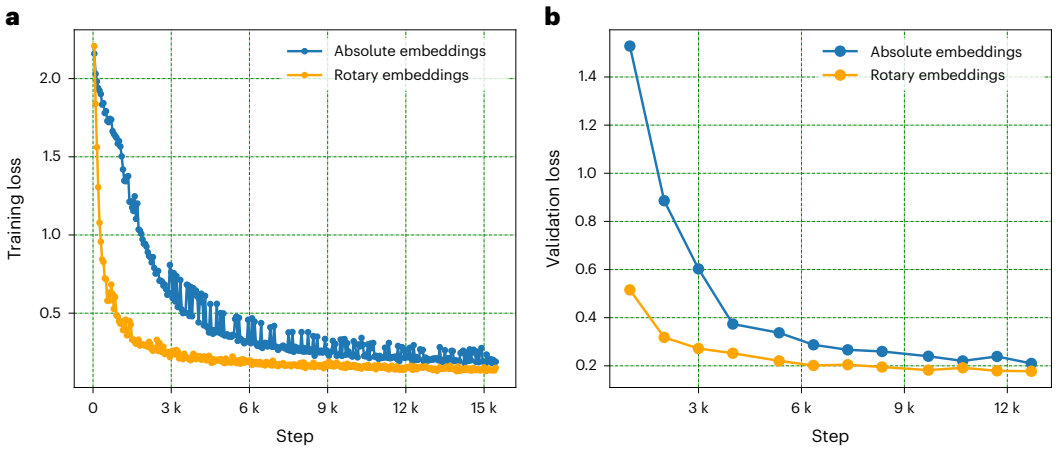

Fig. 2 | Comparison of training and validation losses for absolute and rotary embeddings. a,**b**, Training (**a**) and validation (**b**) losses of our linear attention MoLFormer with rotary (relative) and absolute position embeddings on PubChem. We see that both rotary and absolute MoLFormer have graceful training curves. Our rotary linear attention MoLFormer leads to lower training and validation losses than MoLFormer with absolute position embeddings.

The present study explores the representational power of pretrained chemical language models in 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 accurate three-dimensional (3D) molecular geometric information, which is considered privileged information and not available in general.

Results and discussion

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 Fig. 1. To do so, the MoLFormer model is developed using the masked language model framework[29,30], which randomly masks a certain percentage of tokens within a SMILES sequence during training and then predicts these tokens. The masked language modelling thus exploits self-supervision and enables contextual learning. To allow better contextual learning and faster training, rotary positional embedding[27] was used instead of absolute positional embedding, along with linear attention[31] (see Methods and Supplementary Information for further details of model architecture and training). We saw increased stability and faster convergence in training loss behaviour when pretraining using rotary embeddings, compared with absolute embeddings, as observed in Fig. 2. To demonstrate the effectiveness of the pretrained MoLFormer as a universal and task-agnostic molecular representation, we benchmarked its adaptation performance on numerous challenging classification and regression tasks from MoleculeNet[28]. Details of the benchmark datasets can be found in Supplementary 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 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 Supplementary Table 1. For the fine-tuned strategy, we use a two-layer fully connected network with a hidden dimension of 768 (matching the encoder embedding) with dropout (set to 0.1) and Gaussian error linear unit 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[28], as discussed below. We refer to MoLFormer that has been pretrained on the entire training set comprised of ≈1.1 billion molecules (all molecules from both PubChem and ZINC) as MoLFormer-XL. Unless stated otherwise, MoLFormer-XL is trained with linear attention using rotary positional embeddings and the performance reported is that of the model fine-tuned on the downstream task (see Methods for details). 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 (Supplementary Section C).

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 ML models trained on molecular fingerprints (RF and SVM in Table 1) and graph neural nets. Among the pretrained/self-supervised baselines, Hu et al.[32] pretrain a graph isomorphism network (a GNN that uses a multilayer perceptron and weighted sum of node features in the aggregation) on molecular graphs that includes edge features involved in aggregation. The *N*-gram graph[33] 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[26] is a self-supervised learning framework based on a graph isomorphism network, which uses contrastive loss[34,35]. GraphMVP-C is the graph multiview pretraining framework proposed in ref. 36, where self-supervised learning is performed by leveraging the correspondence and consistency between two-dimensional topological structures and 3D geometric views. We have considered three other geometry-aware GNN baselines, one supervised (DimeNet[37]), and two self-supervised (GeomGCL[36] and GEM[38]). ChemBERTa[25] is a pretrained 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

https://doi.org/10.1038/s42256-022-00580-7

Table 1 | Comparison of fine-tuned MoLFormer with existing supervised and pretrained/self-supervised baselines on multiple classification benchmarks

	BBBP 1	Tox21 12	ClinTox 2	HIV 1	BACE 1	SIDER 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[56]	**85.0**	70.7	63.4	73.8	73.4	55.2
D-MPNN[57]	71.2	68.9	**90.5**	75.0	85.3	63.2
DimeNet[37]	—	78.0	76.0	—	—	61.5
Hu et al.[32]	70.8	78.7	78.9	80.2	85.9	65.2
N-gram[33]	91.2	76.9	85.5	**83.0**	87.6	63.2
MolCLR[26]	73.6	79.8	93.2	80.6	**89.0**	68.0
GraphMVP-C[36]	72.4	74.4	77.5	77.0	81.2	63.9
GeomGCL[36]	—	**85.0**	91.9	—	—	64.8
GEM[38]	72.4	78.1	90.1	80.6	85.6	67.2
ChemBERTa[25]	64.3	—	90.6	62.2	—	—
MoLFormer-XL	**93.7**	84.7	**94.8**	82.2	88.21	**69.0**

Bold indicates the top-performing model. All models were evaluated using the area under the receiver operating characteristic curve on scaffold splits. Baseline performances are adopted from refs. 25, 26,36, '—' signifies that the values were not reported for the corresponding task.

Table 2 | Performance of fine-tuned MoLFormer and other supervised GNN baselines on QM9, QM8, ESOL, FreeSolv and Lipophilicity regression benchmarks

	QM9	QM8	ESOL	FreeSolv	Lipophilicity
GC	4.3536	0.0148	0.970	1.40	0.655
A-FP	2.6355	0.0282	0.5030	0.736	0.578
MPNN	3.1898	0.0143	0.58	1.150	0.7190
MoLFormer-XL	**1.5894**	**0.0102**	**0.2787**	**0.2308**	**0.5289**

For QM9 and QM8, we report average MAE, while root-mean-square error is reported for the remaining tasks. Baseline performances are taken from refs. 28,40. Bold indicates the top-performing model.

scaffold data splits. MoLFormer-XL outperforms all baselines in three (BBBP, ClinTox and SIDER) out of six benchmarks and comes a close second in the other three (Tox21, 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 (see also Supplementary Sections D and E). In particular, QM9 and QM8 involve prediction of 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 ref. 28 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 neighbours before the linear transformation)[39], the attentive FP (A-FP) model[40] and an MPNN variant[18] 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. Supplementary Table 7 further shows MoLFormer outperforming geometry-aware GNNs (DimeNet, GeomGCL and GEM) on three physical property regression benchmarks. These results, combined with MoLFormer-XL performance on the classification benchmarks, confirm its generalizability.

A closer look at QM9. Supplementary Table 9 further compares MoLFormer-XL performance on the QM9 atomization energies and enthalpy (internal energy/enthalpy corrected for reference atomic energy, in electronvolts) prediction tasks with two exemplary supervised 3D GNNs, SchNet[41] and Dimenet[37]. MoLFormer-XL trained on SMILES alone is outperformed by both these models in all of the four tasks. However, SchNet and DimeNet, which directly encode 3D information with specialized architecture for modelling quantum interactions, beat MoLFormer-XL only by roughly a factor of 8 and by roughly a factor of 10, respectively. This result, along with Tables 1 and 2, reinstates the power of learning a universal molecular representation from readily available information, such as SMILES, at a broader scale, while confirming the crucial role of privileged geometric information for quantum-chemical energy prediction. Further, results from this comparison open the door for future investigations with the goal of estimating emergence of geometric awareness in MoLFormer (see later sections) or how the expressiveness of SMILES-only MoLFormer can be further enhanced by adding partial or complete 3D geometric information.

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 into the following three categories: (1) the effect of size and the nature of the pretraining data and model depth, (2) the results without (frozen) and with (fine-tuned) model fine-tuning on the downstream data and (3) the effect of absolute and rotary positional embeddings.

Data/model size. First we investigate how pretraining dataset size affects the performance of MoLFormer-XL on several downstream tasks from the MoleculeNet benchmark. To accomplish this we chose three different weighted combinations of the PubChem and ZINC datasets, specifically a set consisting of 10% of ZINC and 10% of PubChem, another with 100% of PubChem mixed with 10% of ZINC, and then one with 100% of ZINC molecules and 0% of PubChem. We also investigate the influence of model depth by pretraining a six-layer model, named MoLFormer-Base, on the complete ZINC and PubChem dataset. All models are pretrained with rotary embeddings and linear attention and then compared with MoLFormer-XL. Identical learning rates, data splits, optimization and so on are used for pretraining and fine-tuning. Extended Data Tables 1 and 2 summarize these results. While MoLFormer-XL performs better on average, we report two interesting observations. The first is that the model that is pretrained on the second biggest dataset, 100% ZINC, consistently performs worse than all other pretrained models. A possible explanation for the poor performance of the model trained on only ZINC is that the ZINC dataset consists of a much smaller vocabulary than all other dataset combinations, as well as much shorter molecules with little variance with respect to molecule length. 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 and FreeSolv benchmarks in Table 2). Extended Data Tables 1 and 2 further show that MoLFormer-Base has a weaker performance than MoLFormer-XL in the majority of tasks, implying that a deeper model helps in learning.

Fine-tuned versus frozen. Extended Data Table 3 further summarizes the two remaining ablation experiments using the QM9 benchmark. For simplicity we observe that the fine-tuned ablation experiments achieve such a convincing win over the frozen experiments on all pretraining dataset sizes that we opted to only investigate fine-tuning for all other benchmarks. These results provide empirical insights into the neural and data scaling behaviour of MoLFormer.

Table 3 | 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 7,806 molecules from the QM9 test set

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**

Short, medium and long distance categories are defined with interatomic distances in the range of ≤2, 2–4 and 4–10 Å, respectively. Bold indicates the top-performing model.

Position embeddings. The positional embedding ablation results are collected in Extended Data Table 3, which show that MoLFormer with rotary embeddings and fine-tuning is behind the absolute positional embedding model for the smaller datasets, but then wins as the dataset size passes 1 billion molecules.

Insights into MoLFormer

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 Extended Data Table 4 and show that MoLFormer-XL embeddings are better correlated with known molecule similarity measures when compared with ChemBERTa. These results are suggestive of MoLFormer embeddings being informative about 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 within a molecule from the QM9 test set. Spatial distances are obtained from the corresponding energy-minimized geometries provided within the QM9 benchmark[28]. 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 mean absolute error (MAE ≥ 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 (≤2 Å; mostly reflective of typical covalent bonds in the molecule, the C–C single-bond distance being 1.5 Å), medium (2–4 Å) and long (≥4 Å)—and summarize them in Table 3. 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 neighbours in the SMILES sequence. The observed attentions in MoLFormer-XL are slightly more in line with medium and long-range distances, when compared with MoLFormer with full attention. This observation suggests that MoLFormer-XL, with linear attention, does in fact capture spatial relations between atoms more effectively.

Figure 3 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 with the corresponding covalent bond connectivity and 3D distances between atom pairs (complete attention matrices for the same molecules across all layers are shown in Supplementary Figs. 5 and 6). 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 great extent. This capability probably stems from pretraining 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 modelling[42,43]. This is confirmation that structural and diverse property information emerges in the representation learned by a chemical language model pretrained on large-scale data.

Conclusion

In this work, we have explored the power of unsupervised large-scale pretrained molecular language models in 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. This work validates the power of large-scale self-supervised pretrained molecular language models in 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 is indeed aware of interatomic relations within a molecule, even beyond the two-dimensional topology. Finally, at the large-scale learning end, we showcase 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 (1,000 versus 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

https://doi.org/10.1038/s42256-022-00580-7

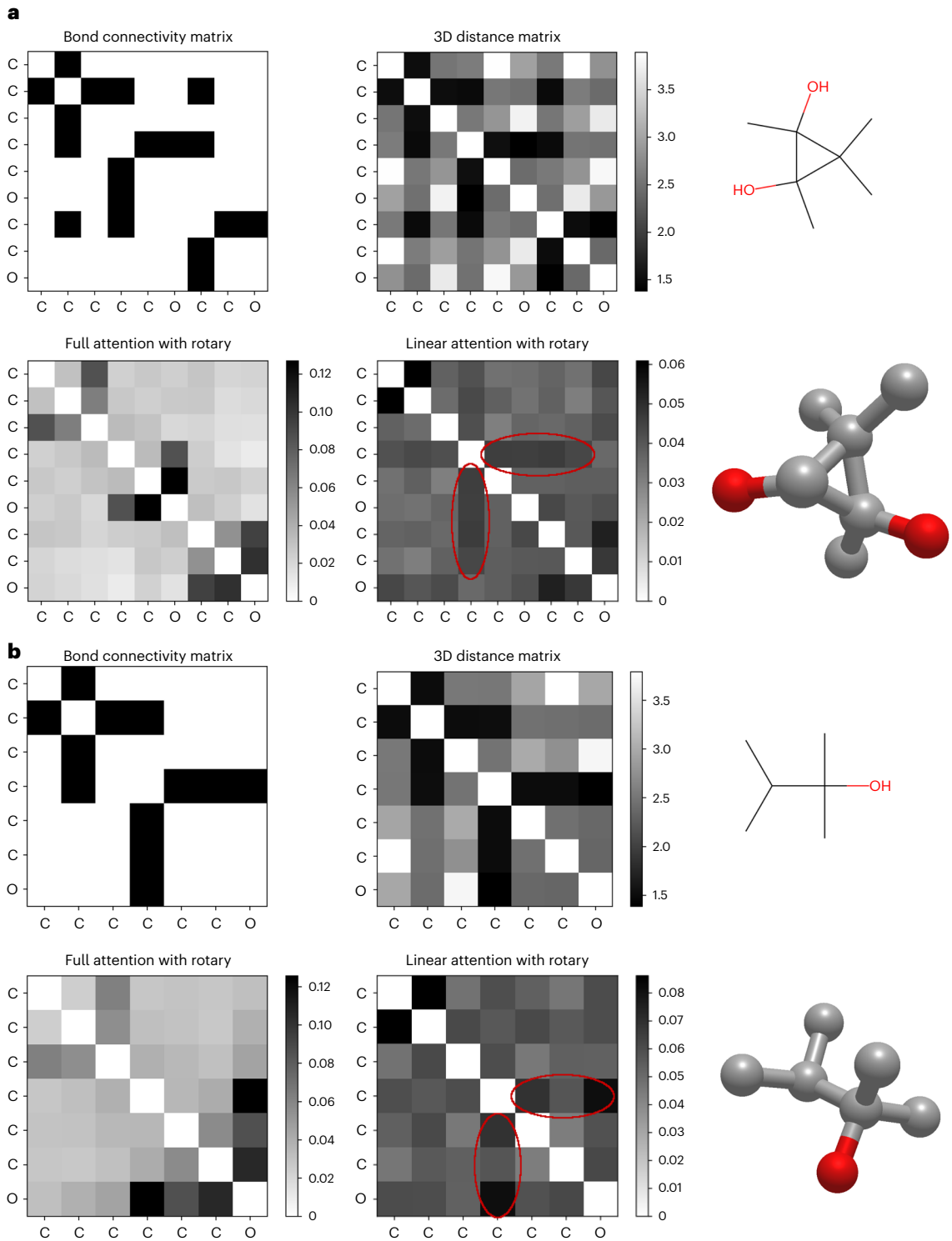

Fig. 3 | a,b, Visualization of the learned attention map (using either full or linear attention) under rotary embedding and corresponding molecular structure (bond connectivity and 3D distance in Angstrom) for two random molecules: 'CC1(C)C(C)(O)C1(C)O' (a) and 'CC(C)C(C)(C)O' (b). The attention map (ranging from 0 to 1; 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 red) the medium-range 3D distance better than does its counterpart.

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[44]. 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 more training data, using improved and/or domain-specific self-supervised tasks, and using other string-based representations such as SELFIES[9].

Methods

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 leverage transformer-based neural nets[23]. Transformers process inputs through a series of blocks alternating between self-attention and feedforward connections. They 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 m as follows:

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

where Q, K and V are the query, key and value respectively. A well known computational bottleneck of the vanilla transformer[23] architecture is that the attention mechanism suffers from a quadratic computational cost with respect to the sequence length. Linear complexity attention models[31,45] have tackled this issue utilizing kernel approximations and random feature approximation variants. This led us to design MoLFormer, which utilizes an encoder based on a transformer with linear attention[31]. MoLFormer with linear attention consists of 12 layers and 12 attention heads per layer, and has a hidden state size of 768. A generalized feature map[31] for the linear attention was chosen (see Supplementary Section A.1.1 for details).

As mentioned above, in a transformer architecture the dependence between tokens at different positions of a (chemical) sequence is modelled under the supervision of position encoding. The seminal work of ref. 23 investigated absolute position embeddings to encode the position of a token in the sequence. More recent work[46–48] showed that use of relative position embeddings between tokens results in improved performance. Rotary position embeddings were introduced in RoFormer[27] as a means to enhance the relative encoding via position-dependent rotations R_m of the query and the keys at m. These rotations can be efficiently implemented as pointwise multiplications and do not result in a marked computational increase.

To leverage rotary embeddings with linear transformers, the use of the following approximation was proposed in ref. 27:

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

where ϕ is a random feature map.

After preliminary experimentation with this linear RoFormer, we found that 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:

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

When compared with ref. 27 we rotate the original keys and queries instead of the transformed ones with ϕ.

Datasets and tokenization

We constructed several datasets for pretraining by combining the PubChem[49] and ZINC[50] datasets with varying proportions 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 ref. 51. All molecules from both PubChem and ZINC are converted to a canonical format utilizing RDKit (http://www.rdkit.org) then tokenized. All unique tokens extracted from the resulting output give us a vocabulary of 2,357 tokens plus 5 special tokens, resulting in a total of 2,362 vocabulary tokens, which are used for all pretrained models considered in this paper, irrespective of pretraining dataset size. In other words, all models have the same embedding capacity with a fixed vocabulary size. However, the unique tokens on which they are pretrained might only contain a subset of the model vocabulary capacity. The post-tokenization sequence length of the molecules ranges from 1 to just over 2,000 tokens. We decided to restrict the sequence length range from 1 token to 202 tokens, inclusive of special tokens, to reduce computation time. Since over 99.4% of all molecules from our dataset contain fewer than 202 tokens, we hypothesize that the removal of molecules with more than 202 tokens would be of minimal negative impact on pretraining.

Large-scale training and parallelization

For pretraining we use the masked language model method defined in ref. 30. Initially 15% of the tokens are selected for possible denoising. From this selection, 80% of the tokens will be randomly selected and replaced with the [MASK] token, 10% of the tokens will be randomly selected to be replaced with a random token and the remaining 10% of the tokens will be unchanged. Training was performed for four epochs through the entire PubChem + ZINC dataset with a fixed learning rate of 1.6×10^{-4} and a batch size of 1,600 molecules per GPU on a total of 16 GPUs over two 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 by up to a factor of 8.

To scale our training to large datasets (1 billion+ data points), we relied on adaptive bucketing of minibatches by sequence length, as well as parallelization via distributed training (see Supplementary Section A for details). Using linear attention and bucketing allowed us to reduce the number of GPUs needed from roughly 1,000 for quadratic attention with no bucketing to 16 (refs. 52–55).

Data availability

Data for model pretraining and fine-tuning on benchmark tasks are available at https://github.com/IBM/molformer.

Code availability

Python codes for MoLFormer training and fine-tuning, and Python notebooks for MoLFormer attention visualization, as well as instances of pretrained models, are available at https://github.com/IBM/molformer. For other enquiries contact the corresponding authors.

References

1. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **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. Rev. Lett.* **108**, 058301 (2012).
3. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **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. Preprint at https://arxiv.org/abs/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. Preprint at https://arxiv.org/abs/1811.08283 (2018).
7. Shin, B., Park, S., Kang, K. & Ho, J. C. Self-attention based molecule representation for predicting drug–target interaction. *Proc. Mach. Learn. Res.* **106**, 230–248 (2019).
8. Daylight Chemical Information Systems *SMARTS—a Language for Describing Molecular Patterns* https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (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 (2020).
10. Gao, W., Fu, T., Sun, J. & Coley, C. W. Sample efficiency matters: a benchmark for practical molecular optimization. In *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022).
11. 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).
12. Duvenaud, D. et al. Convolutional networks on graphs for learning molecular fingerprints. In *NIPS'15: Proc. 28th International Conference on Neural Information Processing Systems* Vol. 2 (2015).
13. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Adv. Neural Inf. Process. Syst.* **29** (2016).
14. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations* (OpenReview.net, 2017).
15. Li, Y., Tarlow, D., Brockschmidt, M. & Zemel, R. Gated graph sequence neural networks. In *4th International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) (OpenReview.net, 2016).
16. Veličković, P. et al. Graph attention networks. In *International Conference on Learning Representations* (2018).
17. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst.* **30**, 1025–1035 (Curran Associates Inc., 2017).
18. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. *Proc. Mach. Learn Res.* **70**, 1263–1272 (2017).
19. Schlichtkrull, M. et al. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference* 593–607 (Springer, 2018).
20. Liao, R., Zhao, Z., Urtasun, R. & Zemel, R. S. LanczosNet: multi-scale deep graph convolutional networks. In *7th International Conference on Learning Representations* (OpenReview.net, 2019).
21. Chen, P., Liu, W., Hsieh, C.-Y., Chen, G. & Zhang, S. Utilizing edge features in graph neural networks via variational information maximization. Preprint at https://arxiv.org/abs/1906.05488 (2019).
22. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823–824 (2004).
23. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 6000–6010 (Curran Associates Inc., 2017).
24. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://doi.org/10.48550/ARXIV.2108.07258 (2021).
25. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. Preprint at https://arxiv.org/abs/2010.09885 (2020).
26. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
27. Su, J., Lu, Y., Pan, S., Wen, B. & Liu, Y. RoFormer: enhanced transformer with rotary position embedding. Preprint at https://arxiv.org/abs/2104.09864 (2021).
28. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
29. Liu, Y. et al. RoBERTa: a robustly optimized BERT pretraining approach. Preprint at https://arxiv.org/abs/1907.11692 (2019).
30. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the NAACL: HLT* Vol 1, 4171–4186 (Association for Computational Linguistics, 2019).
31. Katharopoulos, A., Vyas, A., Pappas, N. & Fleuret, F. Transformers are RNNs: fast autoregressive transformers with linear attention. *Proc. Mach. Learn. Res.* **119**, 5156–5165 (2020).
32. Hu, W. et al. Strategies for pre-training graph neural networks. In *8th International Conference on Learning Representations* (OpenReview.net, 2020).
33. Liu, S., Demirel, M. F. & Liang, Y. *N*-gram graph: simple unsupervised representation for graphs, with applications to molecules. In *NIPS'19: Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. et al.) 8464–8476 (Curran Associates, Inc., 2019).
34. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *Proc. Mach. Learn. Res.* **119**, 1597–1607 (2020).
35. Oord, A. V. D., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://arxiv.org/abs/1807.03748 (2018).
36. Liu, S. et al. Pre-training molecular graph representation with 3D geometry. In *International Conference on Learning Representations* (2022).
37. Klicpera, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. In *8th International Conference on Learning Representations* (OpenReview.net, 2020).
38. Fang, X. et al. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **4**, 127–134 (2022).
39. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **3**, 283–293 (2017).
40. Xiong, Z. et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **63**, 8749–8760 (2019).
41. Schütt, K. et al. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* 992–1002 (Curran Associates Inc., 2017).
42. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA* **118**, e2016239118 (2021).
43. Vig, J. et al. BERTology meets biology: interpreting attention in protein language models. In *9th International Conference on Learning Representations* (OpenReview.net, 2021).
44. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nat. Mach. Intell.* **4**, 189–191 (2022).
45. Choromanski, K. et al. Rethinking attention with Performers. In *Proc. 9th International Conference on Learning Representations* (OpenReview.net, 2021).
46. Shaw, P., Uszkoreit, J. & Vaswani, A. Self-attention with relative position representations. In *Proc. NAACL-HLT* 464–468 (Association for Computational Linguistics, 2018).
47. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).

48. Ke, G., He, D. & Liu, T.-Y. Rethinking positional encoding in language pre-training. In *9th International Conference on Learning Representations* (OpenReview.net, 2021).
49. Kim, S. et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2018).
50. 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).
51. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
52. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: the long-document transformer. Preprint at https://arxiv.org/abs/2004.05150 (2020).
53. Kitaev, N., Kaiser, L. & Levskaya, A. Reformer: the efficient transformer. In *8th International Conference on Learning Representations* (OpenReview.net, 2020).
54. Wang, S., Li, B. Z., Khabsa, M., Fang, H. & Ma, H. Linformer: self-attention with linear complexity. Preprint at https://arxiv.org/abs/2006.04768 (2020).
55. You, Y. et al. Large batch optimization for deep learning: training BERT in 76 minutes. In *8th International Conference on Learning Representations* (OpenReview.net, 2020).
56. Lu, C. et al. Molecular property prediction: a multilevel quantum interactions modeling perspective. *Proc. AAAI* **33**, 1052–1060 (2019).
57. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

Acknowledgement

We thank IBM Research for supporting this work.

Author contributions

All authors conceived the project, developed the MoLFormer framework and designed experiments. J.R., B.B., V.C. and I.P. performed model training, fine-tuning and inference experiments. I.P. and P.D. performed attention map analyses. All authors analysed the results and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-022-00580-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00580-7.

Correspondence and requests for materials should be addressed to Jerret Ross or Payel Das.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-022-00580-7

Extended Data Table 1 | Comparison of MOLFORMER-XL with fine-tuned MOLFORMER models that are either of smaller size or pretrained on smaller datasets on BBBP, HIV, Sider, Clintox, Tox21 and BACE classification benchmarks

Dataset	BBBP	HIV	BACE	SIDER	Clintox	Tox21
10% ZINC + 10% PubChem	91.5	81.3	86.6	68.9	94.6	84.5
10% ZINC + 100% PubChem	92.2	79.2	86.3	69.0	94.7	84.5
100% ZINC	89.9	78.4	87.7	66.8	82.2	83.2
MoLFormer-Base	90.9	77.7	82.8	64.8	61.3	43.2
MoLFormer-XL	**93.7**	**82.2**	**88.2**	**69.0**	**94.8**	**84.7**

https://doi.org/10.1038/s42256-022-00580-7

Extended Data Table 2 | Performance comparison of fine-tuned MOLFORMER-XL with fine-tuned MOLFORMER models are either of smaller size or pretrained on smaller datasets on QM9 (avg MAE), QM8 (avg MAE), ESOL (RMSE), FreeSolv (RMSE), and Lipophilicity (RMSE) regression benchmarks

Dataset	QM9	QM8	ESOL	FreeSolv	Lipophilicity
10% Zinc + 10% Pub	1.7754	0.0108	0.3295	0.2221	0.5472
10% Zinc + 100% Pub	1.9093	**0.0102**	**0.2775**	**0.2050**	0.5331
100% Zinc	1.9403	0.0124	0.3023	0.2981	0.5440
MoLFormer-Base	2.2500	0.0111	0.2798	0.2596	0.6492
MoLFormer-XL	**1.5984**	**0.0102**	0.2787	0.2308	**0.5298**

https://doi.org/10.1038/s42256-022-00580-7

Extended Data Table 3 | Comparison of different MOLFORMER variants on QM9 test set, in terms of average MAE and average standard MAE. Variants considered are MOLFORMER pretrained 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 Supplementary Table 8) is chosen based on the average MAE (Mean Absolute Error) score, lower is better

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**

https://doi.org/10.1038/s42256-022-00580-7

Extended Data Table 4 | Correlation with structural similarity metrics on 10000 randomly selected 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

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