Pre-training Transformers for Molecular Property Prediction Using Reaction Prediction

Johan Broberg [1] **Maria Bånkestad** [1 2] **Erik Ylipää** [1]

Abstract

Molecular property prediction is essential in chemistry, especially for drug discovery applications. However, available molecular property data is often limited, encouraging the transfer of information from related data. Transfer learning has had a tremendous impact in fields like Computer Vision and Natural Language Processing signaling for its potential in molecular property prediction. We present a pre-training procedure for molecular representation learning using reaction data and use it to pre-train a SMILES Transformer. We fine-tune and evaluate the pre-trained model on 12 molecular property prediction tasks from MoleculeNet within physical chemistry, biophysics, and physiology and show a statistically significant positive effect on 5 of the 12 tasks compared to a non-pre-trained baseline model.

1. Introduction

Molecular property prediction has long been used to quickly screen new molecule leads in drug development. The accuracy of these methods is crucial, since false negatives incur high costs when a lead is taken to an experimental phase. Lately, machine learning has become one of the standard tools for molecular property prediction. However, the main challenge is the limited amount of available data to train models on. One solution to this problem is to curate larger datasets using domain expertise. This can be a costly and time consuming approach but has the advantage that larger parts of the relevant molecular domain can be covered (Thawani et al., 2020). Another approach to solve the data scarcity problem is that of *transfer learning* (Zhuang et al., 2019), where knowledge in one task is used to improve performance in another. This is an active area of research in the chemistry domain, mainly using Graph Neural Networks (GNNs) (Kipf & Welling, 2016) and Transformer models (Vaswani et al., 2017). While transfer learning has proven to be very successful in domains such as Natural Language Processing (Howard & Ruder, 2018; Devlin et al., 2018; Liu et al., 2019) and Computer Vision (Girshick et al., 2013), the same clear success is yet to happen in chemistry.

This work is a continuation of the master's thesis (Broberg, 2022) and explores a pre-training strategy for molecular representation learning based on chemical reaction prediction. We use it to pre-train a transformer encoder and compare its performance to a randomly initialized one on a wide range of molecular property prediction tasks. We show statistically significant improvements on 5 of the 12 datasets using a significance level $\alpha = 0.05$ with Bonferroni correction (Bonferroni, 1935; 1936; Noble, 2009).

2. Related Work

Most recent work on pre-training deep models for molecular property prediction uses either GNNs or Transformers.

With GNNs, it is common to use multiple learning objectives that aims to improve representation on different levels (node/edge/graph) (Hu et al., 2020; Rong et al., 2020; Liu et al., 2021). Node and edge level pre-training tasks generally aim to capture graph structural regularities of molecules. Examples of such tasks are prediction of masked node or edge attributes or using node embeddings to predict information about the neighborhood structure. Graph level tasks may also be based on graph structural information but there are also approaches that more explicitly utilize information from the chemistry domain. For example, 3D molecular structure data (Stärk et al., 2021; Fang et al., 2022; Liu et al., 2021) and graph motifs with their connections to functional groups (Rong et al., 2020; Zhang et al., 2021) have been used for graph level pre-training.

For Transformers applied to molecular property prediction using SMILES (Weininger, 1988), a common pre-training approach is to randomly mask parts of the input string. ChemBERTa (Chithrananda et al., 2020), MolBERT (Fabian et al., 2020), and SMILES-BERT (Wang et al., 2019) are

[1]Research Institute of Sweden (RISE), Isafjordsgatan 22, 164 40 Kista, Sweden [2]Department of Information Technology, Uppsala University, Uppsala, Sweden. Correspondence to: Johan Broberg <johan.broberg@ri.se>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

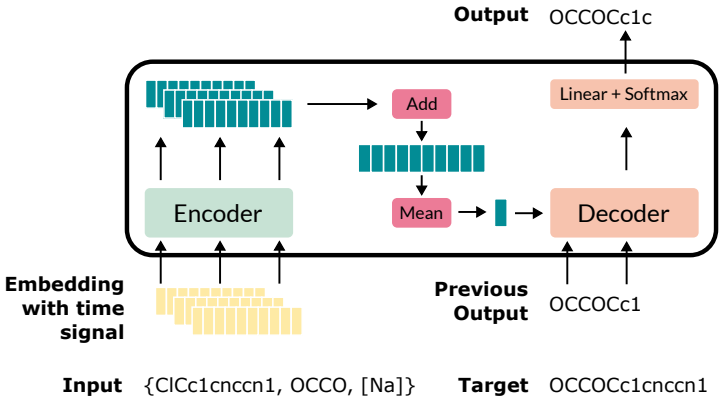

Figure 1. Illustration of our pre-training architecture. Note that each molecule fragment is encoded independently of the others.

only some of the works that explore this method. Though proven beneficial, this approach has not led to the huge improvements seen in NLP by, e.g., the BERT model (Devlin et al., 2018) using the very related masked-language-modeling (MLM) pre-training task. Chithrananda et al. (2020) show a diminishing performance gain on three tasks (BBBP, ClinTox (CT_TOX), and Tox21 (SR-p53)) using a masked token prediction pre-training strategy where an increase in dataset size from 10^6 to 10^7 only lead to a $\approx 3\%$ mean ROC-AUC increase and a $\approx 2\%$ mean PRC-AUC increase. This indicates that pre-training by recovering masked tokens alone might not scale well enough to train powerful property prediction models.

Another recent line of research tries to adapt the Transformer architecture to take molecular graphs as input (Maziarka et al., 2020; Yoo et al., 2020; Ying et al., 2021). Such modifications allow Transformers explicit access to the graph structure, which otherwise must be learned implicitly from string representations. Furthermore, it can also allow information such as node features and edge features to be included in the model input. These models have achieved remarkable results and have in some cases been pre-trained using structure-based tasks (Maziarka et al., 2020; Yoo et al., 2020), but the main contribution to their predictive power seem to stem from their architecture rather than their choice of pre-training task.

A relatively unexplored source of information for pre-training in the chemistry domain is reaction data. To our knowledge, the only published work using such data for pre-training is by Wang et al. (2021). They base their approach on GNNs and a contrastive learning task that teaches the model to encode molecules such that two aggregated sets of molecular encoding vectors lie near each other in the encoding space if they make up the left- and right-hand side of a chemical reaction respectively, and far away if they do not.

Our approach differs from that of Wang et al. (2021) in that we model the pre-training step as a generative reaction prediction task instead of a contrastive learning task. Furthermore, we use a transformer architecture and SMILES representations of molecules while they use a GNN architecture and graph representations.

3. Background

3.1. Chemical Reactions

A chemical reaction is a transformation of one set of molecules into another. The molecules present before the reaction are called *reactants*, while the molecules created through the reaction are called *products*. Molecules might be part of the reaction but not contribute themselves with any atoms to the product molecules created. These are called *reagents*. In reaction prediction one tries to predict the product molecules of a reaction given the reactants and reagents. Normally, a reaction produces multiple product molecules. In our work, we have limited the scope by using data with only a single product.

3.2. SMILES

SMILES, Simplified-Molecular-Input-Line-Entry-System is a linearization of the molecular graph. A molecule has many possible SMILES depending on where the linearization starts and what branches to take. There are certain rules for producing *canonical* SMILES of molecules, where the

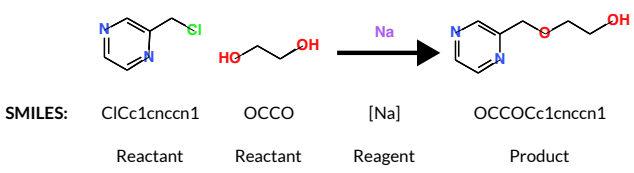

Figure 2. Example reaction with molecules represented as their structural formulas and as SMILES.

same molecule will always encode to the same SMILES. A commonly used strategy when using SMILES as inputs to neural networks is to *randomize* the SMILES, by essentially starting the linearization at a random place and take random branches when traversing the molecular graph. This has shown to act as a powerful data augmentation method when working with SMILES representations of molecules (Bjerrum, 2017; Arús-Pous et al., 2019).

3.3. Transformer

The Transformer is an architecture operating on mathematical sets and was introduced by Vaswani et al. (2017) in the context of neural machine translation. It has been widely used for sequential data such as natural language. A positional encoding is then added to the input data to provide the sequential structure to the model. The original (full) Transformer consists of an encoder and a decoder and is typically used for translation tasks. On its own, the Transformer encoder can be used for sequence classification/regression/representation tasks (Devlin et al., 2018). The key component of the architecture is the multi-head attention mechanism which enables the model to attend to all elements of its input at once. For a detailed description of the full Transformer model and for Transformer encoder models such as BERT we refer to (Vaswani et al., 2017) and (Devlin et al., 2018) respectively.

4. Method

Our approach is based on the Molecular Transformer by Schwaller et al. (2019), in which reaction prediction is modeled as a sequence translation problem for which a full Transformer model is used. Like Schwaller et al. (2019), we represent molecules as SMILES. By using canonicalized SMILES for our product molecules we obtain a fixed target sequence which makes the generative process easier compared to generating graphs, where an order needs to be induced on the edge set (Vinyals et al., 2016).

4.1. Pre-Training Architecture

In the Molecular Transformer, the encoder is applied to SMILES strings containing all reactants and reagents of the corresponding reaction. In our pre-training phase, the Transformer encoder is applied to each reactant and reagent independently. That is, the encoder can only attend to the tokens in the same SMILES fragment, not across fragments. For the set of reactants and reagents $R = \{r_1, r_2, ...\}$ in a given reaction, each SMILES fragment r_i produces an encoding $\mathbf{h}_i \in \mathbb{R}^{L \times d}$ where L is the maximum sequence length and d is the token embedding dimension. The set of such encodings $H = \{\mathbf{h}_1, \mathbf{h}_2, ...\}$ are then aggregated into a single vector $h_R \in \mathbb{R}^d$ representing the entire reaction by first applying element-wise addition across the encodings

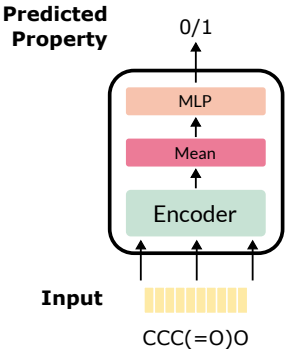

Figure 3. Illustration of the fine-tuning architecture for a binary classification task.

and then averaging across the sequence axis according to

$$h_R = \text{Aggregate}(H) = \text{Mean}(\text{Sum}(H)). \quad (1)$$

The aggregated reaction vector h_R is then passed to the decoders "encoder-decoder attention" layers as a memory key to all tokens in the product (target) SMILES.

We average across the sequence axis because, in the downstream fine-tuning tasks, we will use encoded molecules together with a Multi-Layered Perceptron (MLP). This mean that we will also need to aggregate the encoded molecules into fixed sized vectors. We choose to include this aggregation in the pre-training step, so that the single vector representation will be forced to contain all information needed for the decoding.

4.2. Fine-Tuning Architecture

In the fine-tuning phase, we only use the encoder component from the pre-training phase. Encodings of molecules are aggregated across the sequence axis using the mean. A 2-layered MLP with ReLU activations is used to map the aggregated molecular encodings to the target values. All parameters (encoder and MLP) are tuned in this phase.

4.3. Evaluation

We fine-tune separately on 12 datasets in MoleculeNet (Wu et al., 2017) using 10-folded cross-validation. Each fold (10%) is used once for evaluation of hyperparameter tuning, once for validation, once for testing and otherwise for training. For regression tasks and multi-label classification tasks we use random splits while for single-label classification tasks we use stratified splits. On multi-label prediction benchmarks (PCBA, MUV, TOX21, ToxCast, SIDER, ClinTox) we report the average performance across all tasks as suggested by Wu et al. (2017).

To evaluate the effect of our transfer learning approach, we

Table 1. Performance of the model with and without pre-training showing the mean and standard deviation over a 10-fold cross-validation.

DATA SET	METRIC	MODEL WITHOUT PRE-TRAINING	MODEL WITH PRE-TRAINING	WILCOXON SIGNED-RANK TEST	RANK-BISERIAL CORRELATION
ESOL	RMSE ↓	0.656 ± 0.082	$\mathbf{0.428} \pm 0.077$	**0.001**	1.000
FREESOLV	RMSE ↓	2.057 ± 0.477	$\mathbf{1.484} \pm 0.413$	**0.002**	0.982
LIPOPHILICITY	RMSE ↓	1.012 ± 0.038	$\mathbf{0.700} \pm 0.035$	**0.001**	1.000
PCBA	PRC-AUC ↑	0.178 ± 0.010	0.175 ± 0.006	0.862	0.309
MUV	PRC-AUC ↑	0.023 ± 0.026	0.025 ± 0.017	0.539	0.491
HIV	ROC-AUC ↑	0.704 ± 0.050	$\mathbf{0.757} \pm 0.052$	**0.002**	0.982
BACE	ROC-AUC ↑	0.726 ± 0.088	$\mathbf{0.817} \pm 0.106$	**0.001**	1.000
BBBP	ROC-AUC ↑	0.902 ± 0.104	0.921 ± 0.101	0.019	0.873
TOX21	ROC-AUC ↑	0.799 ± 0.012	0.792 ± 0.013	0.981	0.145
TOXCAST	ROC-AUC ↑	0.693 ± 0.021	0.701 ± 0.013	0.348	0.582
SIDER	ROC-AUC ↑	0.597 ± 0.016	0.578 ± 0.039	0.920	0.255
CLINTOX	ROC-AUC ↑	0.956 ± 0.0044	0.959 ± 0.038	0.652	0.436

compare the pre-trained model to a randomly initialized one which we train directly on each molecular property dataset. Evaluation is based on the best performing checkpoints with respect to the validation set, for each model run. Our null hypothesis is that reaction prediction pre-training has no effect on molecular property prediction across all chemical space. The null hypothesis is then rejected with 95% level of confidence.

We pair the performances on each test fold and use the Wilcoxon signed-rank test (Wilcoxon, 1945) to determine statistically significant differences between our pre-trained model and the randomly initialized one on each of the 12 datasets. The Wilcoxon signed-rank test is a non-parametric version of the Student's t-test, which does not assume normally distributed data. Since we evaluate using multiple tests, we also make a Bonferroni correction to counteract the *multiple comparison problem* (Noble, 2009). Practically this means that, to evaluate our null hypothesis with a 95% level of confidence (significance level $\alpha = 0.05$) we use a significance level of $\alpha_1 = ... = \alpha_m = \alpha/m$ when we test for an effect on each individual dataset. Here $m = 12$ and denote the number of datasets we test on. This means that we for each dataset we use a significance level of $\alpha_1 = ... = \alpha_{12} = 0.05/12 = 0.00417$.

5. Experiment

5.1. Data and Pre-Processing

The dataset used in the pre-training phase is from the USPTO database (Lowe, 2012) and consists of $902,581$ samples used for training and $50,131$ samples used for validation, based on pre-processing and data splits provided by Schwaller et al. (2017).

For the reactants and reagents we use randomized SMILES while the product molecules, our targets, were kept in canonicalized form. Due to the memory complexity of the Transformer we truncate SMILES in the training data to a maximum sequence length of 157. Of the reactions in the pre-training data, 99.9% contain reactants, reagents and products whose SMILES are all shorter than 157, so this truncation threshold only affects 0.1% of the reactions.

5.2. Experimental Setup

For our pre-training model we used a four-layered encoder and decoder, with eight attention heads and a layer width of 256. We pre-trained our model for 150 epochs, using cross-entropy loss and AdamW optimizer with a batch size of 4096 and a cosine cyclic learning rate scheduler with base learning rate of 10^{-5} and maximum learning rate of $5 \cdot 10^{-4}$.

In the fine-tuning phase we tuned the learning rate for all models on each fold in the cross validation. We did this in order to fairly compare the performance. Each tuning was based on 20 runs with learning rates sampled geometrically in the interval $\left[10^{-6}, 10^{-3}\right]$. During learning rate tuning we trained each run for 50 epochs for all datasets except PCBA and MUV which were only tuned for one respectively ten epochs due to the large number of samples in these datasets. The batch size was set to 64. For the number of epochs, we used early stopping with 40 update steps of patience. Throughout this work we tokenized SMILES by converting each character to their corresponding ASCII-value.

5.3. Results

The results from our experiment are shown in table 1. For each dataset we present the mean and standard deviation over the 10-folded cross-validation along with the p-value of the Wilcoxon signed-rank test and the rank-biserial cor-

relation. On 5 of the 12 downstream property prediction tasks our pre-training strategy show a statistically significant positive effect given the Bonferroni-corrected significance level $\alpha = 0.00417$. We therefore reject our null hypothesis and conclude that our pre-training approach using reaction prediction has an effect on molecule property prediction

6. Limitations

This pre-training strategy has two limitations that we would like to point out. We use reactions that only have one major product molecule, but most reactions contain more than one product. This is an advantage of the approach proposed by Wang et al. (2021). Another limitation is that we base our statistical analysis on cross-validation of the downstream tasks. A more robust statistical analysis would have been based on several different runs of pre-training.

7. Conclusions

In this paper we have presented a pre-training strategy for transformer models using reaction prediction. We demonstrate a statistically significant effect on 5 out of 12 datasets from MoleculeNet and conclude that reaction prediction pre-training has an effect on molecular property prediction. Our results show, in line with Wang et al. (2021), that chemical reactions can be used to successfully pre-train models for downstream molecular property prediction tasks.

References

Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., Chen, H., and Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *Journal of cheminformatics*, 11(1):1–13, 2019.

Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. *CoRR*, abs/1703.07076, 2017. URL http://arxiv.org/abs/1703.07076.

Bonferroni, C. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936. URL https://books.google.se/books?id=3CY-HQAACAAJ.

Bonferroni, C. E. Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, pp. 13–60, 1935.

Broberg, J. Pre-training molecular transformers through reaction prediction. Master's thesis, KTH, 2022.

Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885, 2020. URL https://arxiv.org/abs/2010.09885.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Fabian, B., Edlich, T., Gaspar, H., Segler, M. H. S., Meyers, J., Fiscato, M., and Ahmed, M. Molecular representation learning with language models and domain-relevant auxiliary tasks. *CoRR*, abs/2011.13230, 2020. URL https://arxiv.org/abs/2011.13230.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2): 127–134, Feb 2022. ISSN 2522-5839. doi: 10.1038/s42256-021-00438-4. URL https://doi.org/10.1038/s42256-021-00438-4.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. URL http://arxiv.org/abs/1311.2524.

Howard, J. and Ruder, S. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146, 2018. URL http://arxiv.org/abs/1801.06146.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for Pre-training Graph Neural Networks. *arXiv:1905.12265 [cs, stat]*, February 2020. URL http://arxiv.org/abs/1905.12265. arXiv: 1905.12265.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. URL http://arxiv.org/abs/1609.02907.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3D geometry. *CoRR*, abs/2110.07728, 2021. URL https://arxiv.org/abs/2110.07728.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Lowe, D. M. *Extraction of chemical structures and reactions from the literature*. Dissertation, University of Cambridge, 2012.

Maziarka, L., Danel, T., Mucha, S., Rataj, K., Tabor, J., and Jastrzebski, S. Molecule attention transformer. *CoRR*, abs/2002.08264, 2020. URL https://arxiv.org/abs/2002.08264.

Noble, W. S. How does multiple testing correction work? when prioritizing hits from a high-throughput experiment, it is important to correct for random events that falsely appear significant. how is this done and what methods should be used? *Nature Biotechnology*, 27:1135+, December 2009.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data, 2020. URL https://arxiv.org/abs/2007.02835.

Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C., and Laino, T. "found in translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Computing Research Repository*, abs/1711.04810, 2017. URL http://arxiv.org/abs/1711.04810.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, Aug 2019. ISSN 2374-7951. doi: 10.1021/acscentsci.9b00576. URL http://dx.doi.org/10.1021/acscentsci.9b00576.

Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., and Liò, P. 3D infomax improves gnns for molecular property prediction. *CoRR*, abs/2110.04126, 2021. URL https://arxiv.org/abs/2110.04126.

Thawani, A. R., Griffiths, R.-R., Jamasb, A., Bourached, A., Jones, P., McCorkindale, W., Aldrick, A. A., and Lee, A. A. The photoswitch dataset: A molecular machine learning benchmark for the advancement of synthetic chemistry, 2020. URL https://arxiv.org/abs/2008.03226.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Vinyals, O., Bengio, S., and Kudlur, M. Order matters: Sequence to sequence for sets. 11 2016.

Wang, H., Li, W., Jin, X., Cho, K., Ji, H., Han, J., and Burke, M. D. Chemical-reaction-aware molecule representation learning. *CoRR*, abs/2109.09888, 2021. URL https://arxiv.org/abs/2109.09888.

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pp. 429–436, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366663. doi: 10.1145/3307339.3342186. URL https://doi.org/10.1145/3307339.3342186.

Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, feb 1988. ISSN 0095-2338. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. S. MoleculeNet: A benchmark for molecular machine learning. *Computing Research Repository*, abs/1703.00564, 2017. URL http://arxiv.org/abs/1703.00564.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T. Do transformers really perform bad for graph representation? *CoRR*, abs/2106.05234, 2021. URL https://arxiv.org/abs/2106.05234.

Yoo, S., Kim, Y., Lee, K. H., Jeong, K., Choi, J., Lee, H., and Choi, Y. S. Graph-aware transformer: Is attention all graphs need? *CoRR*, abs/2006.05213, 2020. URL https://arxiv.org/abs/2006.05213.

Zhang, Z., Liu, Q., Wang, H., Lu, C., and Lee, C.-K. Motif-based graph self-supervised learning for molecular property prediction, 2021. URL https://arxiv.org/abs/2110.00987.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685, 2019. URL http://arxiv.org/abs/1911.02685.