nature portfolio

https://doi.org/10.1038/s42256-022-00447-x

Supplementary information

Molecular contrastive learning of representations via graph neural networks

Supplementary Information

Molecular Contrastive Learning of Representations via Graph Neural Networks

Yuyang Wang[1,2], Jianren Wang[3], Zhonglin Cao[1], and Amir Barati Farimani[1,2,4,*]

[1]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[4]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[*]corresponding author: Amir Barati Farimani (barati@cmu.edu)

A Details of Molecular Datasets

Table 1 summarizes all the benchmarks used in our work. These benchmarks from MoleculeNet[1] cover a wide variety of molecular properties, including physiology (i.e., BBBP, Tox21, SIDER, ClinTox), biophysics (i.e., BACE, MUV, HIV), physical chemistry (i.e., FreeSolv, Lipo, ESOL), and quantum mechanics (i.e., QM7, QM8, QM9). Also, numbers of data vary significantly among the benchmarks, ranging from less than 1K to more than 130K. All benchmarks except QM9 are scaffold split to train/validation/test sets by the ratio of 8/1/1, which provides a more challenging yet realistic setting. Random splitting is implemented on QM9 following the settings in most related works[2–4] for comparison. ROC-AUC is used as the metric for classification tasks while RMSE and MAE are used for regression tasks.

Dataset	# Molecules	# Tasks	Task type	Metric	Split
BBBP	2039	1	Classification	ROC-AUC	Scaffold
Tox21	7831	12	Classification	ROC-AUC	Scaffold
ClinTox	1478	2	Classification	ROC-AUC	Scaffold
HIV	41127	1	Classification	ROC-AUC	Scaffold
BACE	1513	1	Classification	ROC-AUC	Scaffold
SIDER	1427	27	Classification	ROC-AUC	Scaffold
MUV	93087	17	Classification	ROC-AUC	Scaffold
FreeSolv	642	1	Regression	RMSE	Scaffold
ESOL	1128	1	Regression	RMSE	Scaffold
Lipo	4200	1	Regression	RMSE	Scaffold
QM7	6830	1	Regression	MAE	Scaffold
QM8	21786	12	Regression	MAE	Scaffold
QM9	130829	8	Regression	MAE	Random

Table 1. Summary of all the benchmarks for molecular property predictions used in this work.

We follow Hu et al.[5] to build a simple yet unambiguous set of node and bond features to embed the two-dimensional (2D) molecular graph. RDKit is used to convert SMILES to a 2D graph and extract the features. The details of node and edge features can be found in Table 2. When a node is masked, the atomic number is set to 119 and chirality to unspecified.

Feature type	Feature name	Range
Node feature	Atomic number Chirality	[1, 119] {unspecified, tetrahedral CW, tetrahedral CCW, other}
Edge feature	Bond type Bond direction	{single, double, triple, aromatic} {none, end-upright, end-downright}

Table 2. Node and edge features used in MolCLR.

B Detailed Results of QM9

Table 3 reports detailed results on QM9 database. The property name, unit, mean and std of test MAE for all the models are included. Not surprisingly, SchNet[2] and MGCN[3] outperform the other models greatly. These two models successfully develop interaction layers, which elaborately take quantum interactions into consideration as titles of both works indicate. Besides, both models include 3D positional information as the input, which benefits quantum mechanics property predictions. However, MolCLR pre-training is still demonstrated to be effective on this challenging benchmark. MolCLR shows better prediction accuracy in 7 out 8 tasks among all the pre-training/self-supervised models. $\text{MolCLR}_{\text{GIN}}$ surpasses Hu et al.[5] in all the tasks, which also utilizes GIN as the encoder. Besides, in comparison to GCN and GIN trained via supervised learning, $\text{MolCLR}_{\text{GCN}}$ and $\text{MolCLR}_{\text{GIN}}$ improve the performance on all the tasks within QM9. MolCLR also obtains lower test MAE when set side-by-side with another supervised baseline, D-MPNN[6].

Property	ε_{HOMO}	ε_{LUMO}	$\Delta\varepsilon$	ZPVE	μ	α	$\langle R^2 \rangle$	C_v
Unit	eV	eV	eV	eV	D	bohr^3	bohr^2	cal/mol K
RF	0.186±0.001	0.276±0.002	0.269±0.001	0.276±0.000	0.658±0.004	3.245±0.015	121.837±0.124	1.738±0.003
SVM	0.148±0.000	0.234±0.002	0.248±0.004	0.157±0.000	0.750±0.004	4.065±0.057	189.510±1.078	1.795±0.010
GCN[7]	0.115±0.010	0.133±0.007	0.174±0.013	0.075±0.018	0.532±0.015	1.495±0.338	43.325±15.140	0.514±0.209
GIN[8]	0.097±0.005	0.103±0.010	0.138±0.004	0.055±0.021	0.483±0.004	1.315±0.405	35.278±6.779	0.457±0.073
SchNet[2]	0.041±0.001	0.034±0.003	0.063±0.002	0.002±0.000	0.033±0.001	0.235±0.061	0.073±0.002	0.033±0.000
MGCN[3]	0.042±0.001	0.057±0.002	0.064±0.001	0.001±0.000	0.056±0.002	0.030±0.007	0.113±0.001	0.038±0.001
D-MPNN[6]	0.093±0.005	0.106±0.002	0.148±0.003	0.037±0.004	0.450±0.006	0.493±0.008	24.371±0.922	0.244±0.005
HU. et.al[5]	0.116±0.000	0.118±0.000	0.161±0.001	0.083±0.001	0.543±0.001	1.725±0.008	55.418±0.291	0.705±0.012
N-Gram[4]	0.142±0.001	0.138±0.001	0.193±0.001	0.009±0.000	0.540±0.002	0.611±0.022	59.137±0.178	0.334±0.007
$\text{MolCLR}_{\text{GCN}}$	0.104±0.000	0.110±0.001	0.149±0.001	0.045±0.004	0.507±0.002	0.644±0.053	26.600±0.257	0.259±0.011
$\text{MolCLR}_{\text{GIN}}$	0.087±0.000	0.092±0.000	0.127±0.000	0.033±0.004	0.464±0.001	0.463±0.017	17.425±0.919	0.164±0.002

Table 3. Test MAE of different models for each property in QM9.

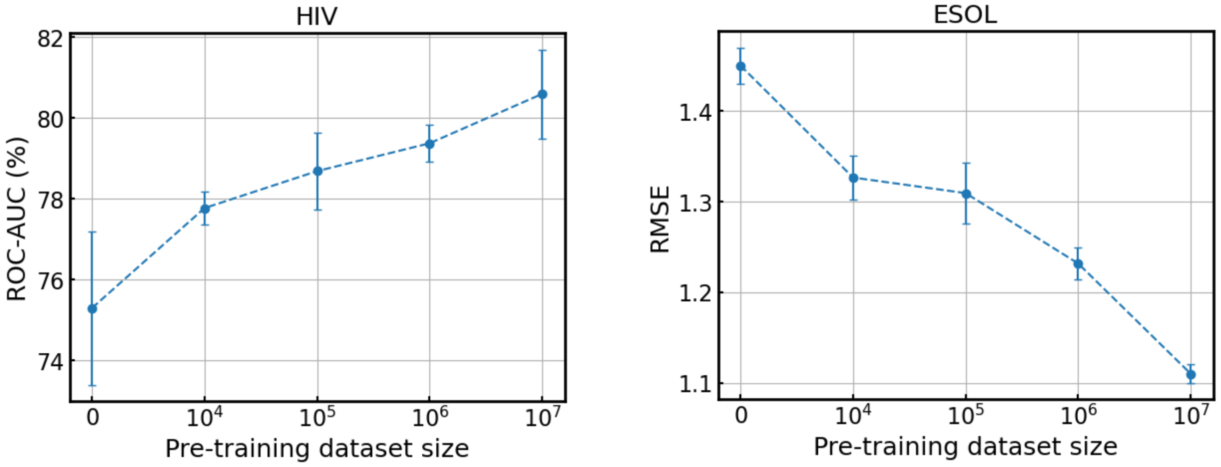
C Investigation of Pre-training Datasets for MolCLR

MolCLR pre-training makes use of the large unlabeled molecular data. We also investigate whether pre-training on certain dataset benefits molecular property predictions of its own. To this end, we conduct MolCLR pre-training on MUV and QM9 as shown in Table 4, since these are the two largest datasets in MoleculeNet[1]. Within the table, MolCLR$_{\text{PubChem}}$ denotes MolCLR framework pre-trained on the ∼10M unlabeled molecules from PubChem[9]. MolCLR$_{\text{MUV}}$ and MolCLR$_{\text{QM9}}$ indicates pre-training on MUV and QM9, respectively. The training and fine-tuning follow the same setting reported in the main manuscript. To avoid data leakage, we split the MUV and QM9 into train/validation/test by the ratio of 8:1:1 and only pre-train the models on the training splits. When conducting fine-tuning on MUV, MolCLR pre-training on MUV improves the test ROC-AUC by 15.4% and MolCLR pre-trained on QM9 also obtains a great improvement by 14.9% in comparison with no pre-training. Not surprisingly, MolCLR pre-trained on the 10M dataset performs the best better as it benefits from larger unlabeled molecular data. Similarly, on QM9, pre-training on MUV and QM9 decreases the test MAE by 1.553 and 2.010, respectively. As expected, MolCLR$_{\text{PubChem}}$ achieves the larges improvement by 2.384 on QM9. Therefore, pre-training on the dataset itself via MolCLR boosts the performance significantly. Also, the pre-trained model on one dataset can be directly transferred to another and outperforms training from scratch.

	Metric	Supervised	MolCLR$_{\text{MUV}}$	MolCLR$_{\text{QM9}}$	MolCLR$_{\text{PubChem}}$
MUV	ROC-AUC (%)	71.8±2.5	87.2±2.1	86.7±2.8	88.6±2.2
QM9	MAE	4.741±0.912	3.188±0.441	2.731±0.019	2.357±0.118

Table 4. Comparison of MolCLR pre-training on different datasets. Test ROC-AUC (%) are reported for MUV and MAE for QM9. Supervised indicates supervised learning with no pre-training. MolCLR$_{\text{MUV}}$ MolCLR$_{\text{QM9}}$, and MolCLR$_{\text{PubChem}}$ denote MolCLR pre-training on the MUV, QM9, and ∼10M PubChem, respectively.

We further probe the influence of the magnitude of pre-training datasets on MolCLR. Subsets of size 10K, 100K, and 1M are randomly sampled from the whole ∼10M PubChem pre-training dataset. Figure 1(a) and Figure 1(b) report the test results of different pre-training data size on HIV and ESOL databases. Pre-training dataset size 0 indicates supervised learning is directly conducted without pre-training. As the number of data increases, the averaged test HIV ROC-AUC increases from 75.3 to 80.6. Similarly, the larger the dataset, the lower test RMSE on ESOL is observed. Also, even pre-training on a small dataset, i.e., 10K molecules, GNN models gain obvious improvements in comparison to supervised learning. For example, pre-training on 10K data improves ROC-AUC by 2.5% on HIV and decreases RMSE by 0.12 on ESOL. It is demonstrated that MolCLR benefits from the large dataset, and therefore can be widely used for the huge unlabeled molecule data. On the other hand, MolCLR pre-training on a small dataset still boosts the performance compared to supervised learning, which demonstrates the effectiveness of the contrastive learning framework on molecule graphs.

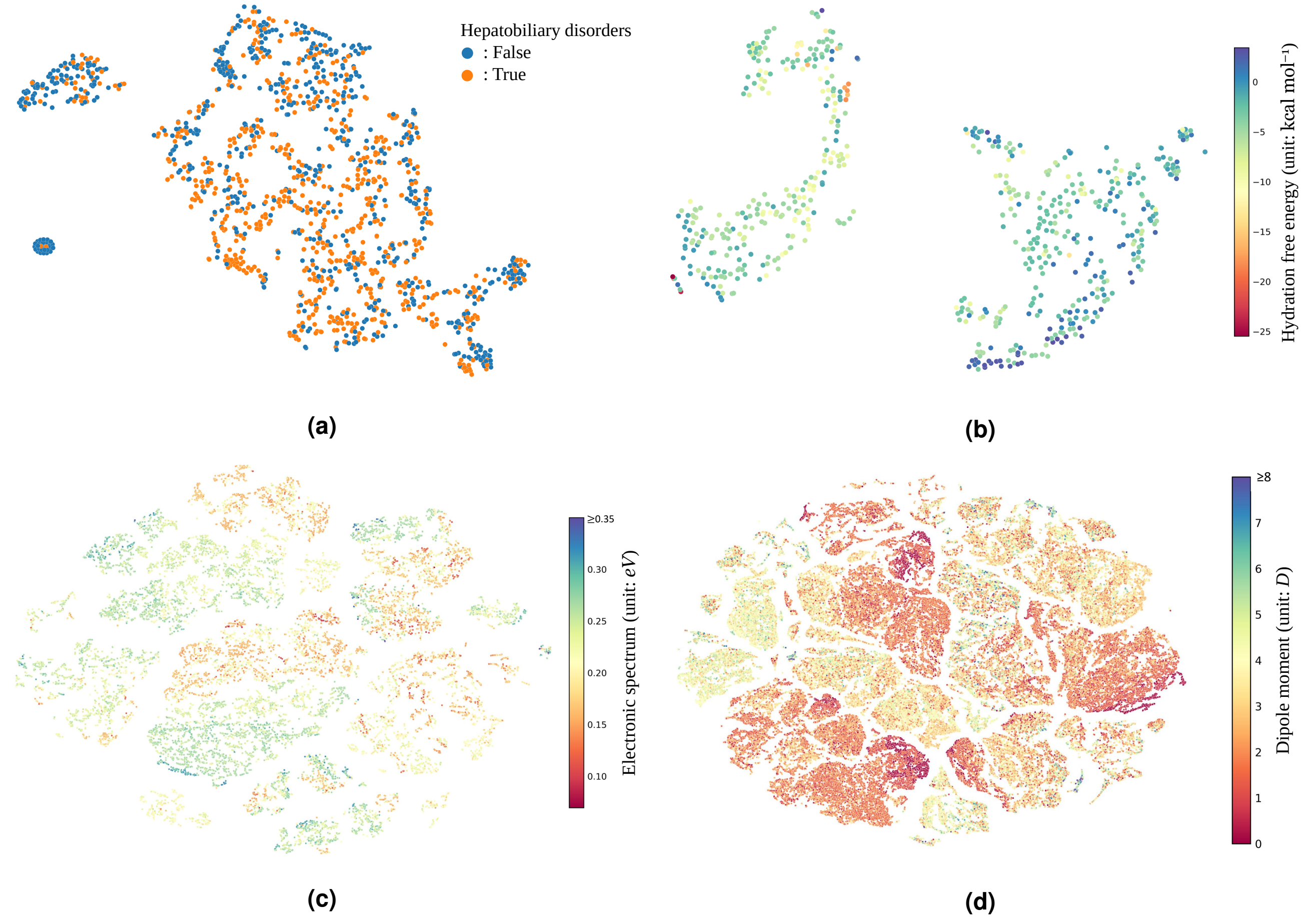


(a) **(b)**

Figure 1. Results of MolCLR pre-training on different dataset sizes. (a) Test ROC-AUC (%) on HIV. (b) Test RMSE on ESOL.

D Visualization of MolCLR Representations

Besides, to illustrate the representations from the pre-trained MolCLR, we visualize the molecule features via t-SNE, where molecules are from various databases and colored by corresponding property labels (Figure 2). Notice that all the features are extracted directly from pre-trained MolCLR without fine-tuning. Namely, the model has no access to the molecular property labels during training. Figure 2 shows molecules from SIDER[10], FreeSolv[11], QM8[12,13], QM9[14]. Features from pre-trained MolCLR show clustering based on the labels, even without accessing labels during training. For instance, in Figure 2(d), molecules are colored by the dipole moment μ. Molecules with relatively high μ (green and blue) are clustered on the bottom right, whereas molecules with low μ (dark red) are clustered in the center of the plot. Similar clustering trends can also be observed in other t-SNE visualizations in Figure 2.



(a)

(b)

(c)

(d)

Figure 2. Two-dimensional t-SNE embedding of the molecular representations learned by our MolCLR pre-training. (a) Molecules from SIDER database and color indicates whether the molecule causes hepatobiliary disorder side effect. (b) Molecules from FreeSolv database and color indicates hydration free energy of each molecule. (c) Molecules from QM8 database and color indicates the electronic spectrum calculated from CC2 of each molecule. (d) Molecules from QM9 database and color indicates the averaged electronic spectrum μ of each molecule.

E More Results of Molecule Retrieval via MolCLR

In this section, more examples of molecule retrieval based on MolCLR-learned representations are shown in Figure 3. Nine molecules that are closest to the query molecule in the MolCLR representation domain are listed with RDKFP and ECFP similarities labeled. Notably, molecules with close MolCLR representations also have high FP similarities. Also, the selected molecules share similar structures and functional groups. For instance, in Figure 3(a), all listed molecules share functional groups like sulfonyl groups and nitrogen heterocycles. Also, in Figure 3(b), the first molecule at the second row is exactly the same as the query molecule except for few carbon-carbon bonds. These examples further demonstrate that through contrastive learning, MolCLR automatically learns chemically meaningful representations.

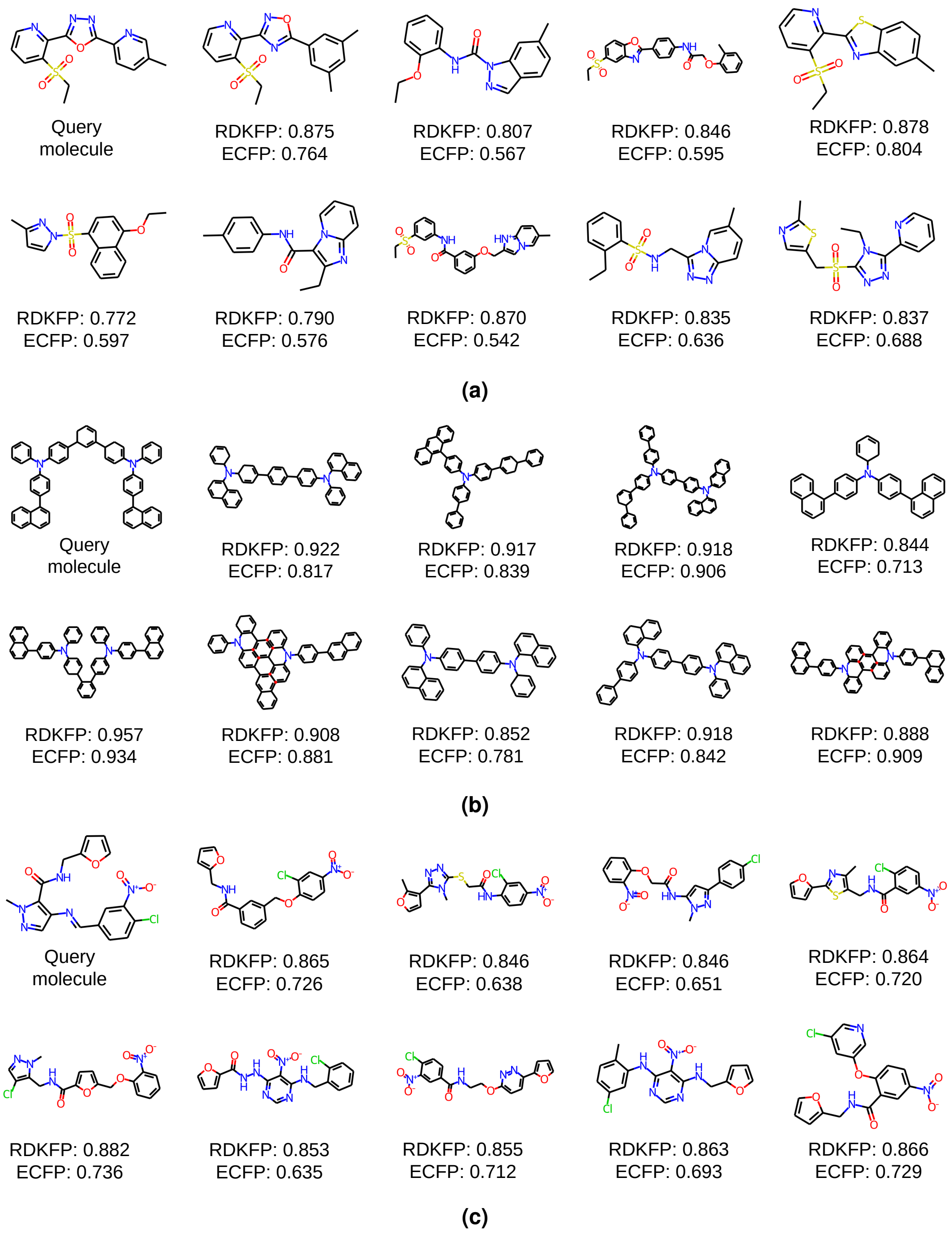


Figure 3. Three Query molecules (PubChem ID (a) 130187714 (b) 132175476 (c) 4862714) and 9 closest molecules for each query molecule in MolCLR representation domain with RDKFP and ECFP similarities labeled.

F Temperature in Contrastive Loss

The choice of the temperature parameter τ in NT-Xent loss[15] impacts the performance of contrastive learning[15]. An appropriate τ benefits the model to learn from hard negative samples. To investigate τ for molecule representation learning, we train MolCLR with three different temperatures: 0.05, 0.1, and 0.5 as shown in Table 5. We report the averaged ROC-AUC (%) over all the seven classification benchmarks using 25% subgraph removal as the augmentation strategy. It is demonstrated that $\tau = 0.1$ performs the best in the downstream molecular tasks. Therefore, we use $\tau = 0.1$ as the temperature in the following experiments.

Temperature (τ)	0.05	0.1	0.5
ROC-AUC (%)	76.8±1.2	80.2±1.3	78.4±1.7

Table 5. Influence of temperature τ in NT-Xent loss for MolCLR. Mean and standard deviation of all the seven classification benchmarks are reported.

G Fine-tuning Details

During fine-tuning for each downstream task, we randomly search the hyper-parameters to find the best performing setting on the validation set and report the results on the test set. Table 6 lists the combinations of different hyper-parameters. Besides, we also consider if cosine annealing learning rate decay[16] improves the fine-tuning performance. In addition, we randomly pick MolCLR-trained GNNs at different epoch as the initialization for fine-tuning.

Name	Description	Range
`batch_size`	Input batch size	$\{32, 128, 256\}$
`lr`	Initial learning rate for MLP head	$\{5 \times 10^{-4}, 10^{-3}\}$
`lr_base`	Initial learning rate for the pre-trained GNN base	$\{5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$
`dropout`	Dropout ratio for the GNN	$\{0, 0.1, 0.3, 0.5\}$
`n_layer`	Number of hidden layers in MLP	$\{1, 2\}$
`hidden_size`	Size of hidden layers in MLP	$\{256\}$
`activation`	Nonlinear activation function in MLP	{ReLU[17], Softplus[18]}

Table 6. Fine-tuning hyper-parameters for pre-trained MolCLR model.

References

1. Wu, Z. *et al.* Moleculenet: a benchmark for molecular machine learning. *Chem. science* **9**, 513–530 (2018).
2. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *The J. Chem. Phys.* **148**, 241722 (2018).
3. Lu, C. *et al.* Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 1052–1060 (2019).
4. Liu, S., Demirel, M. F. & Liang, Y. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. In *NeurIPS* (2019).
5. Hu, W. *et al.* Strategies for pre-training graph neural networks. In *International Conference on Learning Representations* (2020).
6. Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *J. chemical information modeling* **59**, 3370–3388 (2019).
7. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
8. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations* (2019).
9. Kim, S. *et al.* Pubchem 2019 update: improved access to chemical data. *Nucleic acids research* **47**, D1102–D1109 (2019).
10. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The sider database of drugs and side effects. *Nucleic acids research* **44**, D1075–D1079 (2016).
11. Mobley, D. L. & Guthrie, J. P. Freesolv: a database of experimental and calculated hydration free energies, with input files. *J. computer-aided molecular design* **28**, 711–720 (2014).
12. Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *J. chemical information modeling* **52**, 2864–2875 (2012).
13. Ramakrishnan, R., Hartmann, M., Tapavicza, E. & Von Lilienfeld, O. A. Electronic spectra from tddft and machine learning in chemical space. *The J. chemical physics* **143**, 084111 (2015).
14. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. data* **1**, 1–7 (2014).
15. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).
16. Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
17. Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, vol. 30, 3 (Citeseer, 2013).
18. Glorot, X., Bordes, A. & Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323 (JMLR Workshop and Conference Proceedings, 2011).