Supplementary Materials of MolMVC

Zhijian Huang [1], Ziyu Fan [1], Siyuan Shen [1], Min Wu [2], Lei Deng [1,*]

[*]To whom correspondence should be addressed.

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China and
[2]Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 138632, Singapore

contact: zhijianhuang@csu.edu.cn

Contents

1 Details of datasets and methods

1.1 Details of pre-training dataset and downstream datasets

For the pre-training dataset, we leverage the PCQM4Mv2 dataset sourced from the OGB Large-Scale Challenge[1]. This dataset comprises 3.4 million molecular data instances, each accompanied by SMILES [2] and atomic coordinates in molecular 3D space. We utilize RDKit [3] to convert SMILES notations into a 2D graph structure. This includes the extraction of atom features, bond features, and the generation of adjacency matrices that capture the connectivity relationships within the molecular structure. Additionally, we employ ESPF, a method providing information about molecular substructure in a 1D format, to enrich the representation.

For the downstream tasks, we select Molecular Property Prediction (MPP), Cancer Drug Response (CDR), Drug-Target Binding Affinity (DTA), and SARS-CoV-2 drug repositioning. Molecular property, closely tied to clinical efficacy [4], represents a fundamental aspect in drug design. To assess our model, we adopt six classification benchmark datasets sourced from the widely used MoleculeNet [5]. These datasets encompass both biophysical and physiological tasks. Following established practices [5, 6], we employ scaffold splitting [7] to divide each dataset into 8:1:1 for training, validation, and testing, respectively. Each dataset contains a different number of binary tasks. The specifics of the 6 datasets utilized in our study are provided below. The summary can be found in Table S1.

Table S1: Summary for molecular property prediction tasks.

Dataset	Number of tasks	Number of molecules
BBBP	1	2,039
BACE	1	1,513
Clintox	2	1,478
HIV	1	41,127
Sider	27	1,427
ToxCast	617	8,576

1. **BBBP** (Blood–brain Barrier Penetration): Predicts drug penetration into the brain

with a dataset of 2,039 compounds, focusing on blood-brain barrier permeability.

2. **BACE** (β-secretase 1 Inhibitors): Quantifies binding results for 1,522 compounds inhibiting human β-secretase 1, representing IC50 affinity and binary inhibitor class.

3. **ClinTox** (Clinical Toxicology): Compares FDA-approved drugs with those failing clinical trials (1,491 compounds) due to toxicity, examining clinical trial toxicity and FDA approval status.

4. **HIV** (Human Immunodeficiency Virus): Screens 41,127 compounds for inhibiting HIV replication, presenting a classification task between inactive and active compounds.

5. **SIDER** (Side Effect Resource): Database of 1,427 approved drugs, classifying side effects into 27 organ classes, particularly focusing on hepatological disorders.

6. **Toxcast** (Toxicological data collection): Toxicological data (8615 compounds) for an extensive compound library is obtained through in vitro high-throughput screening, involving experiments across more than 617 tasks.

CDR and DTA, representing two drug-related tasks, utilize the GDSC2 [8] for CDR, and Davis [9] and Kiba [10] datasets for DTA. Specifically, to ensure fair comparisons, the same datasets and data segmentation are employed as in DeepTTA [11] and GraphDTA [12] for CDR and DTA tasks, respectively. For SARS-CoV-2 drug repositioning, we utilize the SARS-CoV-2 dataset from DeepCoVDR [13] for fine-tuning. This dataset comprises 318 drugs and their anti-SARS-CoV-2 activity. As a target dataset, we utilize the ReFrame dataset [14], which contains 17 identified ReFRAME actives.

1.2 Experimental settings

In the pre-training stage, the learning rate is 1e-4, the batch size is 800, the masking ratio is 30%, the epoch is 20 and the optimizer is Adam. With the pre-trained MolMVC, we conduct extensive downstream experiments, including MPP, DTA prediction, CDR

prediction, and SARS-CoV-2 drug repositioning. For the molecular property prediction, we fine-tune the pre-trained MolMVC and the experiments configuration can be found in Table S2. The experiments configuration of DTA, CDR and SARS-CoV-2 drug repositioning are same as their state-of-art methods GraphDTA, DeepTTA and DeepCoVDR respectively. All experiments are conducted on 2 NVIDIA RTX4090 GPUs and are repeated three times with different seeds.

Table S2: Hyperparameter and performance summary of molecular property prediction.

Dataset	Learning Rate	Weight Decay	Dropout	Batch Size
BBBP	3×10^{-5}	1×10^{-2}	0.5	80
BACE	2×10^{-5}	1×10^{-2}	0.5	80
Clintox	1×10^{-4}	1×10^{-2}	0.5	80
HIV	8×10^{-5}	1×10^{-7}	0.5	30
Sider	5×10^{-5}	1×10^{-5}	0.5	60
ToxCast	2×10^{-5}	1×10^{-2}	0.5	60

1.3 Input features

For 1D views, the input is ESPF containing 2585 substructures.

For 2D view, the input includes atom features and bond features. The atom features contain atom number(1-118) and chirality(unspecified, tetrahedralcw, tetrahedralccw, other). The bond features contain bond type(single, double, triple, aromatic), bond direction(none, endupright, enddownright) and bond stereo(stereonone, stereoz, stereoe, stereocis, stereotrans, stereoany).

For 3D view, the input includes atom features and position features. We use the same atom feature as 2D view and the position feature is the spatial position of atoms.

1.4 Calculation process of Transformer, GIN and Schnet

1.4.1 Transformer

Transformer mainly includes two parts, a multiattention layer and a feed-forward network [15]. A multi-head attention layer consists of parallel self-attention layers to calculate scaled dot-product attention which can consider different information subspaces of substructures. Denote h_n^{1d} as the input of n-th block, $h_0^{1d} = H_{m,input}^{1d}$ and the multi-head attention algorithm work with u heads as follow:

$$\begin{aligned} \text{head}_i^{1d} &= \text{softmax}\left(\frac{Q_n^{1d}K_n^{1d}}{\sqrt{d}}\right)V_n^{1d} \\ &= \text{softmax}\left(\frac{h_n^{1d}W_n^Q(h_n^{1d}W_n^K)^T}{\sqrt{d}}\right)h_n^{1d}W_n^V, \end{aligned} \tag{1}$$

$$\text{MultiHead}\left(h_n^{1d}\right) = \text{concat}\left(\text{head}_1^{1d}, \ldots, \text{head}_u^{1d}\right)W_n^T, \tag{2}$$

where W_n^Q, W_n^K, W_n^V and W_n^T are weight parameters, and $\frac{1}{\sqrt{d}}$ is a scaling factor depending on the dimension of each head. Then, a feed-forward layer is conducted to allow model have stronger reasoning ability:

$$H_m^{1d} = f\left(\text{MultiHead}\left(h_n^{1d}\right)W_n^{f,1} + b_n^{f,1}\right)W_n^{f,2} + b_n^{f,2}, \tag{3}$$

where $W_n^{f,1}$, $W_n^{f,2}$, $b_n^{f,1}$ and $b_n^{f,2}$ are weight parameters and H_m^{1d} is the output of n^{th} Transformer encoder block.

1.4.2 Graph isomorphism network

The propagation pattern of the l^{th} layer in Graph Isomorphism Network (GIN)[16] can be calculated as follows:

$$h_a^{2d,l+1} = \text{MLP}\left(h_a^{2d,l} + \sum_{b\in\mathcal{N}(a)}\left(h_b^{2d,l} + \text{MLP}\left(h_{ab}^{2d,l}\right)\right)\right), \tag{4}$$

where $h_a^{2d,l}$ and $h_{ab}^{2d,l}$ is the 2D atom representation and neighboring bond representation at l-th layer, MLP is a multi-layer perceptron.

1.4.3 SchNet

The key part of SchNet[17] is the continuous-filter convolution layer:

$$h_a^{3d,t+1} = \sum_{v=1}^{d} h_v^{3d,t} \circ W^t \mathbf{e_k} \left(r_a - r_v \right), \tag{5}$$

where $h_a^{3d,t+1}$ is the 3D atom representation of (t+1)-th layer, W^t is a weight parameter and $\mathbf{e_k}$ is the distance between atom a and v with radial basis functions.

2 Additional experiments and analysis

2.1 Results of drug repositioning

To transfer MolMVC knowledge to specific SARS-CoV-2 drugs, we fine-tune pre-trained MolMVC for 5 epochs and 10 epochs on the SARS-CoV-2 dataset. Subsequently, we compare the drug repositioning ability of MolMVC with COVIDS-2 [18] and DeepCoVDR[13] on the ReFrame dataset, with no overlap between the ReFrame dataset and our fine-tuning dataset. As illustrated in Figure S1, MolMVC with 10 fine-tuning epochs accurately distinguishes ReFrame actives, with predicted scores for all 17 ReFrame actives exceeding 0.8. This underscores the stronger drug repositioning ability of MolMVC compared to the comparative methods, showcasing its capacity to transfer generalized molecular knowledge to a specific domain. Notably, increasing the finetuning epochs from 5 to 10 results in significantly improved predicted scores for ReFrame actives, indicating higher adaptability in SARS-CoV-2 drug tasks as more specific knowledge is incorporated into MolMVC.

To further demonstrate that MolMVC can assist drug repositioning, we used the MolMVC model fine-tuned for 10 epochs on the SARS-CoV-2 dataset to search for potential antiviral COVID-19 drugs among 1853 FDA-approved drugs from [19]. We conducted a nonexhaustive quality search for the top 20 compounds, at least 8 of which have been shown to have potential to treat SARS-CoV-2. The results can be found in the Table S3.

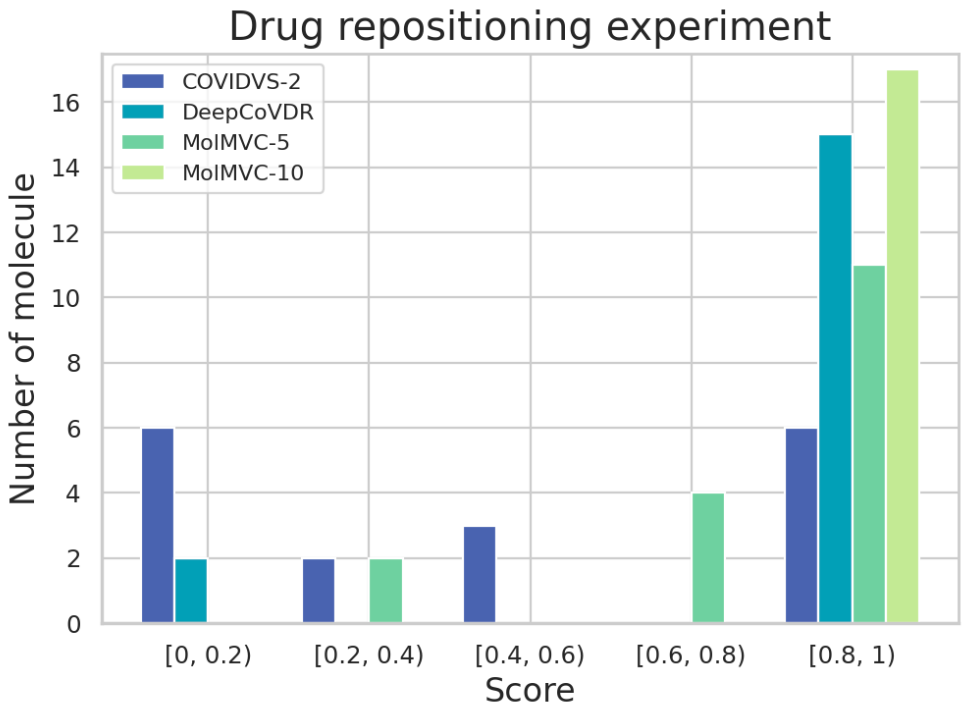

Figure S1: Distribution of scores for 17 ReFRAME actives predicted with MolMVC and comparison methods. MolMVC-5 is the MolMVC finetuned for 5 epochs using the SARS-CoV-2 dataset, while MolMVC-10 is the MolMVC fine-tuned for 10 epochs.

Table S3: Potential antiviral COVID-19 drugs predicted by MolMVC in FDA-approved drugs with the support of literature

Compound name	Ranking position	DOI
Thioguanine	1	https://doi.org/10.1101/2020.07.01.183020
Acyclovir	4	https://doi.org/10.31579/2690-4861/199
Nelarabine	12	https://doi.org/10.1021/acsbiomedchemau.2c00039
Inosine	13	https://doi.org/10.1016/j.jpha.2022.10.002
Papaverine	14	https://doi.org/10.1002/ddr.21961
Kinetin	15	https://doi.org/10.1038/s41467-023-35928-z
Didanosine	16	https://doi.org/10.1021/acsomega.1c07095
Regadenoson	20	https://doi.org/10.1371/journal.pone.0288920

2.2 Distribution and alignment of molecular representations

To further investigate the impact of pre-training strategies on molecular representation, we selected 500 molecules from the validation set of pre-training dataset and calculated the cosine similarity between the MolMVC representations of each pair. The distributions of cosine similarity before and after pre-training are presented in Figure S2(A) and (B). The results show our pre-training scheme ensures that the molecular representation distribution of different modalities obtained by MolMVC is reasonable and reliable.

We also randomly select 1000 molecules from the validation set of pre-training dataset and use centered kernel alignment (CKA)[20] with RBF kernels to compare representations between different layers of encoders for the three modalities before and after pre-training. As shown in Figure S3, we can observe that the addition of the pre-training scheme achieves high-level alignment of representations between different molecular modalities, and this alignment increases with the number of model layers. It is worth mentioning that after pre-training, the alignment effect between the three molecular modalities has been significantly improved, demonstrating that the pre-trained MolMVC exhibits high alignment.

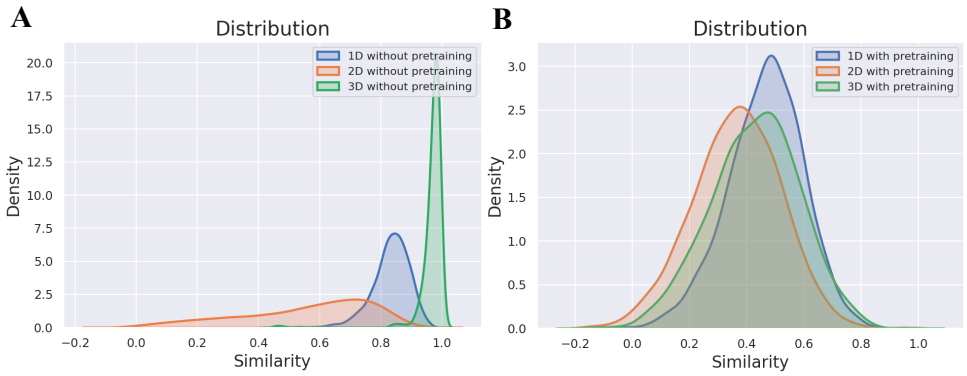

Figure S2: Molecular representation distribution. (A) and (B) shows distribution of MolMVC representation before and after pre-training.

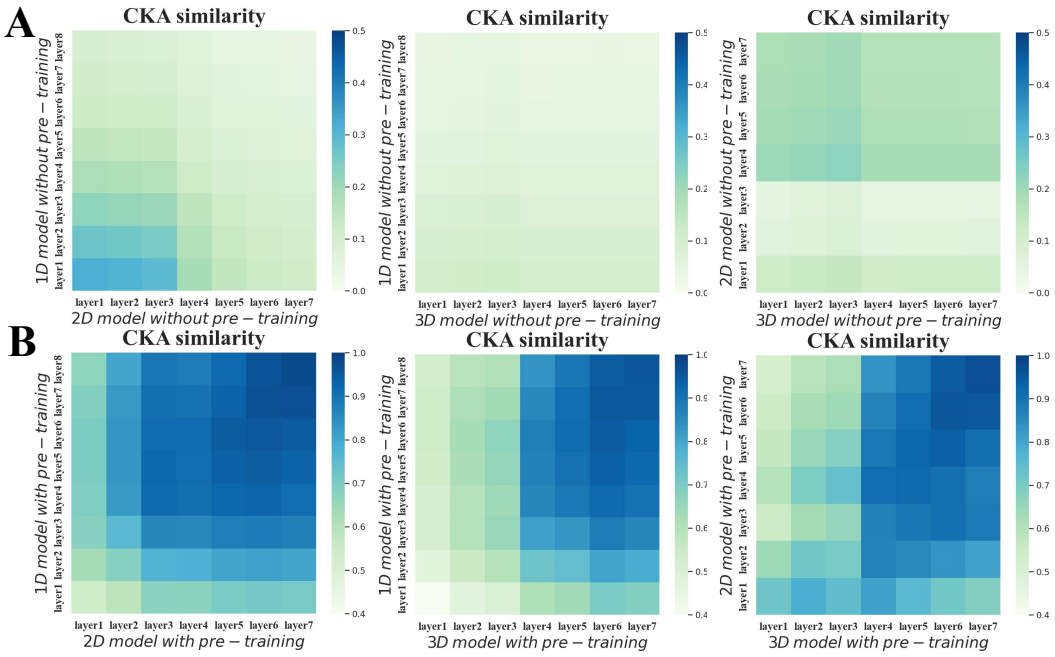

Figure S3: Heatmaps of molecular representation comparison using CKA. (A) and (B) shows the CKA similarity between different layers of encoders for three modalities before and after pre-training.

2.3 More results of investigation of molecular representation

To show more results of investigation of molecular representation, we select BACE dataset for t-SNE visualization and molecular retrieval. The results are shown in Figure 1. As can be observed, the results further demonstrate the informative of our molecular representation and the effectiveness of our pre-traing scheme.

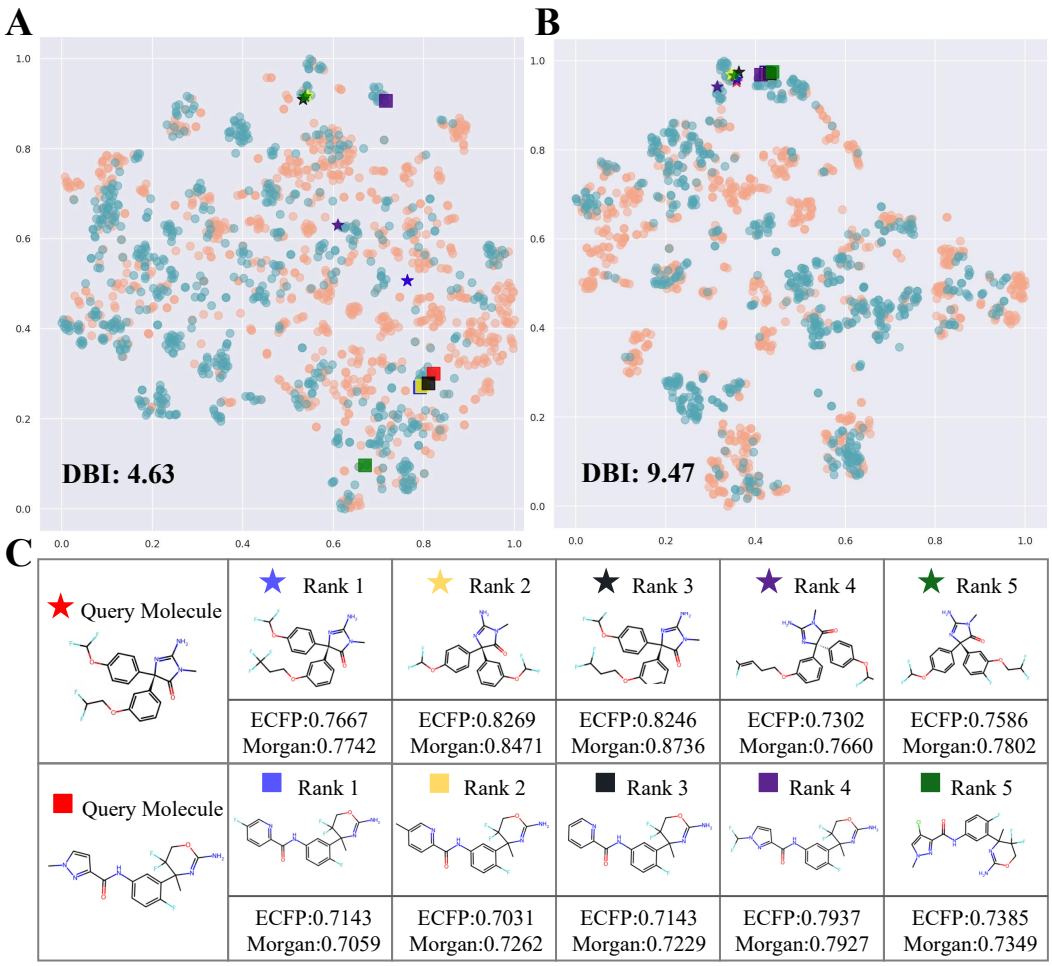

Figure S4: More results of investigation of molecular representation. (A) and (B) are t-SNE visualization of representation of BACE dataset with and without pre-training. (C) is the results of molecular retrieval.

References

[1] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*, 2021.

[2] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[3] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.

[4] Manu De Rycker, Beatriz Baragaña, Suzanne L Duce, and Ian H Gilbert. Challenges and recent progress in drug discovery for tropical diseases. *Nature*, 559(7715):498–506, 2018.

[5] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[6] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Mole-bert: Rethinking pre-training graph neural networks for molecules. 2023.

[7] Bharath Ramsundar, Peter Eastman, Pat Walters, and Vijay Pande. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. " O'Reilly Media, Inc.", 2019.

[8] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.

[9] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

[10] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, 2014.

[11] Likun Jiang, Changzhi Jiang, Xinyu Yu, Rao Fu, Shuting Jin, and Xiangrong Liu. Deeptta: a transformer-based model for predicting cancer drug response. *Briefings in Bioinformatics*, 23(3):bbac100, 2022.

[12] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

[13] Zhijian Huang, Pan Zhang, and Lei Deng. Deepcovdr: deep transfer learning with graph transformer and cross-attention for predicting covid-19 drug response. *Bioinformatics*, 39(Supplement_1):i475–i483, 2023.

[14] Laura Riva, Shuofeng Yuan, Xin Yin, Laura Martin-Sancho, Naoko Matsunaga, Lars Pache, Sebastian Burgstaller-Muehlbacher, Paul D De Jesus, Peter Teriete, Mitchell V Hull, et al. Discovery of sars-cov-2 antiviral drugs through large-scale compound repurposing. *Nature*, 586(7827):113–119, 2020.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[17] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

[18] Shiwei Wang, Qi Sun, Youjun Xu, Jianfeng Pei, and Luhua Lai. A transferable deep learning approach to fast screen potential antiviral drugs against sars-cov-2. *Briefings in Bioinformatics*, 22(6):bbab211, 2021.

[19] Ctibor Skuta, Martin Popr, Tomas Muller, Jindrich Jindrich, Michal Kahle, David Sedlak, Daniel Svozil, and Petr Bartunek. Probes & drugs portal: an interactive, open data resource for chemical biology. *Nature methods*, 14(8):759–760, 2017.

[20] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.