Graph Self-Supervised Learning for Optoelectronic Properties of Organic Semiconductors

Zaixi Zhang [1 2] Qi Liu [1 2] Shengyu Zhang [3] Chang-Yu Hsieh [3] Liang Shi [4] Chee-Kong Lee [5]

Abstract

The search for new high-performance organic semiconducting molecules is challenging due to the vastness of the chemical space, machine learning methods, particularly deep learning models like graph neural networks (GNNs), have shown promising potential to address such challenge. However, practical applications of GNNs for chemistry are often limited by the availability of labelled data. Meanwhile, unlabelled molecular data is abundant and could potentially be utilized to alleviate the scarcity issue of labelled data. Here, we advocate the use of self-supervised learning to improve the performance of GNNs by pre-training them with unlabeled molecular data. We investigate regression problems involving ground and excited state properties, both relevant for optoelectronic properties of organic semiconductors. Additionally, we extend the self-supervised learning strategy to molecules in non-equilibrium configurations which are important for studying the effects of disorder. In all cases, we obtain considerable performance improvement over results without pre-training, in particular when labelled training data is limited, and such improvement is attributed to the capability of self-supervised learning in identifying structural similarity among unlabeled molecules.

1. Introduction

Organic semiconductors (OSCs) have been a vibrant field of research since the discovery of their electroluminescence properties in the 1960s and 1970s (Ostroverkhova, 2016; Christensen et al., 2021b) due to their potential applications in solar cells (Hains et al., 2010; Myers & Xue, 2012; Lu et al., 2015; Hedley et al., 2017), light-emitting devices(Minaev et al., 2014; Xu et al., 2016) and field-effect transistors (Sirringhaus, 2014). The use of organic materials offers several advantages as compared to their inorganic counterparts, such as low production costs, versatile synthesis processes, and high portability. However the search of new high performance OSCs has proved challenging due to the vastness of chemical space. Computational simulation could assist the search for OSCs materials with desirable electronic properties critical to their electronic applications at a lower cost compared to experiments. Despite the efficiency of computational simulations, quantum chemistry methods such as the density functional theory (DFT) are still too expensive for high-throughput virtual screening involving a large number of candidate molecules (Hachmann et al., 2011). Recent successful applications of machine learning (ML) in chemistry show that it could accurately predict various molecular and material properties with vastly higher efficiency compared to quantum chemistry calculations (Christensen et al., 2021a; Behler, 2011; 2015; 2016; Poltavsky & Tkatchenko, 2021; Smith et al., 2018; Taylor et al., 2021; Kulichenko et al., 2021; Dral et al., 2018; Chen et al., 2018; Wang et al., 2021; Dral, 2020; von Lilienfeld et al., 2020; Noé et al., 2020; Olivares-Amaya et al., 2011; Sajeev et al., 2013; Kanal et al., 2013; Shu & Levine, 2015; Li et al., 2015; Pyzer-Knapp et al., 2016; Gómez-Bombarelli et al., 2016; Nagasawa et al., 2018; Jørgensen et al., 2018; Janai et al., 2018; Sahu et al., 2018; Padula & Troisi, 2019; Padula et al., 2019; Lee, 2019; St. John et al., 2019; Lederer et al., 2019; Roch et al., 2020; Simine et al., 2020; Lu et al., 2020; Farahvash et al., 2020; Pereira et al., 2017; Duan et al., 2020; Welborn et al., 2018; Prezhdo, 2020; Musil et al., 2018; Mahapatra et al., 2018; Atahan-Evrenk & Atalay, 2019; Bian et al., 2019; Thawani et al., 2020), these include ML modeling of potential energy (Christensen et al., 2021a; Behler, 2011; 2015; 2016; Poltavsky & Tkatchenko, 2021), OSCs (Olivares-Amaya et al., 2011; Sajeev et al., 2013; Kanal et al., 2013; Shu & Levine, 2015; Li et al., 2015; Pyzer-Knapp et al., 2016; Gómez-Bombarelli et al., 2016; Nagasawa et al., 2018; Jørgensen et al., 2018; Janai et al., 2018; Sahu et al., 2018; Padula & Troisi, 2019; Padula

[1] Anhui Province Key Lab of Big Data Analysis and Application, University of Science and Technology of China [2] State Key Laboratory of Cognitive Intelligence, Hefei, Anhui, China [3] Tencent Quantum Lab [4] Chemistry and Biochemistry, University of California, Merced [5] Tencent America. Correspondence to: Chee-Kong Lee <cheekonglee@tencent.com>.

2nd AI4Science Workshop at the 39^{th} International Conference on Machine Learning (ICML), 2022. Copyright 2022 by the author(s).

et al., 2019; Lee, 2019; St. John et al., 2019; Lederer et al., 2019; Roch et al., 2020; Simine et al., 2020; Lu et al., 2020; Farahvash et al., 2020), non-adiabatic dynamics (Wang et al., 2021; Dral et al., 2018; Chen et al., 2018) and electronic structures (Pereira et al., 2017; Duan et al., 2020; Welborn et al., 2018). In particular, state-of-the-art deep learning methods such as the graph neural networks (GNNs) have shown to be able to achieve prediction accuracy superior to other traditional ML methods (Ramakrishnan et al., 2014; Duvenaud et al., 2015; Schütt et al., 2017b; Gilmer et al., 2017; Wu et al., 2018; Lu et al., 2019; Schütt et al., 2018; 2019; Chen et al., 2019; Unke & Meuwly, 2019; Klicpera et al., 2020; Liu et al., 2020; Qiao et al., 2020; Hao et al., 2020).

Despite its immense potential, practical application of GNNs in chemistry is frequently limited by the availability of labelled training data. Meanwhile, in many cases unlabelled data are abundant, e.g. from publicly available database like PubChem or molecular dynamics simulations. In order to utilize the availability of these unlabelled data and overcome the scarcity of labelled data, recently various self-supervised pre-training strategies have been devised for GNNs, and have been successfully demonstrated in social network and biological domains (Xie et al., 2021; Lu et al., 2021; Hu* et al., 2020; Rong et al., 2020; Zhang et al., 2021). However its applications on quantum mechanical properties have been limited and only available on simple small molecules like those in the QM7 and QM8 datasets (Rong et al., 2020). Additionally, these pre-training strategies have not been tested on excited state properties or molecules in non-equilibrium configurations. In this work we advocate the use of self-supervised learning (SSL) in GNNs for predicting the optoelectronic properties of OSCs. SSL pre-training of GNNs consists of two steps: unsupervised learning and supervised fine-tuning. During the unsupervised learning stage, a GNN is first trained on a large collection of unlabeled molecular data such that it derives generic transferable knowledge encoding the intrinsic graph representation of molecules. During the fine-tuning stage the pre-trained GNN model is fine-tuned on task-specific molecular data, such that it adapts the generic knowledge for specific tasks.

For the first application in this work, we apply SSL to the prediction of optoelectronic properties of organic photovoltaic molecules where the only input is the SMILES strings of the molecules. For the second application, we extend the SSL strategy to molecules in non-equilibrium configurations by incorporating 3D coordinates into the training of SSL. Existing SSL studies only focus on molecules in their equilibrium geometries and thus do not consider the effect of disorder or temperature on the electronic properties of OSC molecules. However the presence of disorder could have significant implication on the performance of OSC devices. For example it is known that the existence of disorder can limit the transport of charges and excitons (Lee et al., 2019), leading to a drop in device efficiency. On the other hand disorder can sometimes assist the separation of charge-transfer exciton, an important step for efficient organic photovoltaics (Deotare et al., 2015; Shi et al., 2017). Thus to design high-performance OSCs, it is essential to understand the effects of disorder on the electronic properties of OSCs. Therefore in this work, we also explore the application of SSL on predicting the excited state properties of OSC molecules in non-equilibrium configurations.

2. Graph neural networks

In contrast with traditional ML methods where hand-crafted molecular descriptors are required as input, deep learning methods such as GNNs are capable of extracting informative representation of a molecule solely from atom types and Cartesian coordinates. In GNNs, the basic chemical information of molecules are encoded as computational graphs and these graphs are used as the input for the graph-based training algorithm. As compare to the traditional ML methods, GNNs are capable of representing the irregular molecular graph structures more naturally. Specifically, a computational graph $G = (V, E)$ is defined as the connectivity relations between a set of nodes (V) and a set of edges (E). Naturally, a molecule can also be considered as a graph consisting of a set of atoms (nodes) and a set of bonds (edges).

A schematic of the GNNs is shown in Fig. 1a. After a molecule is converted into a computational graph, each node (atom) is represented by an embedding vector. GNNs learn the optimal representation of each atom using a message passing algorithm that iteratively aggregates the information of its neighboring atoms and the corresponding edges (Gilmer et al., 2017). After the message passing phase, the molecule-level embedding vectors can be generated by pooling all the atoms through summation. Finally, the learned molecular representation can be used for the prediction of molecular properties through the read-out phase. It is worth noting that alternative pooling strategies can also be used, such as the maximum function or attention-based layers (Wang et al., 2019). Our version of GNNs are implemented using PyTorch (Adam et al., 2017), and the source codes can be found on GitHub (Sou).

3. Self-supervised learning

We use node and edge-level attribute masking for SSL of GNNs in this work: first the node and edge attributes of the graphs are masked, then the GNNs are tasked to predict those attributes based on neighboring structures (Hu* et al., 2020). Fig.1 (b) illustrates the working mechanism of

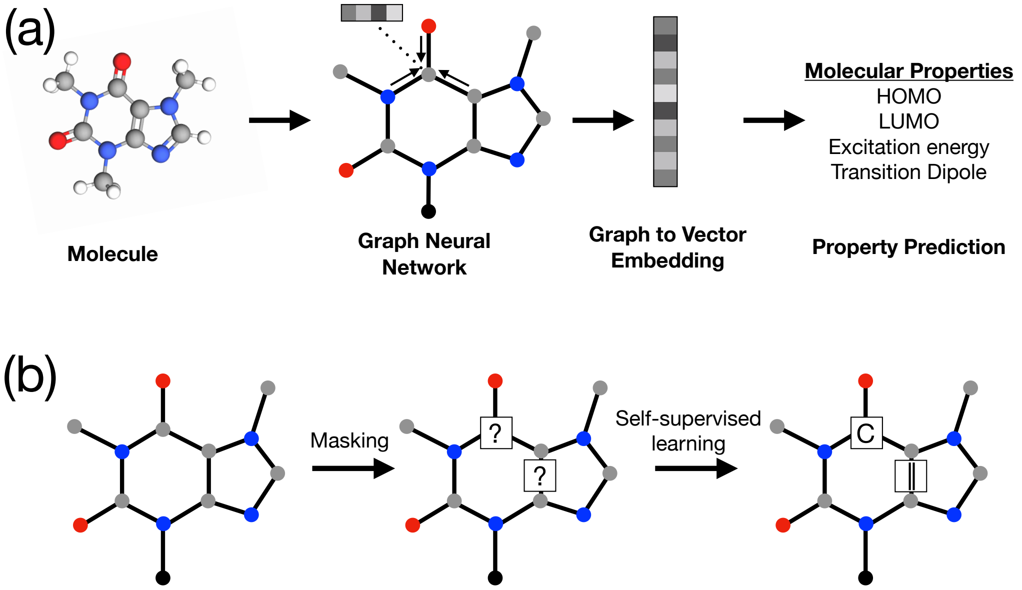

Figure 1. A schematic of the model used in this work. (a) A molecule is first converted into a computational graphs, and each node (i.e. atom) is represented by an embedding vector. The optimal node embedding is learned via a message passing algorithm, i.e. the embedding vector is iteratively updated by aggregating the embeddings of its neighboring nodes and edges. The molecule-level embedding vectors can then be generated by pooling all the atoms through summation. The learned molecular representation can be used for the prediction of molecular properties through the read-out phase. (b) In self-supervised learning (SSL), we randomly mask 15% of the node (i.e atom) and edge (i.e. bond) attributes, and the GNNs are tasked to predict these masked attributes.

attribute masking when applied to a molecular graph. We randomly mask the atom and bond types of the molecular graphs by replacing them with special masked indicators. GNNs are then tasked to predict these masked node or edge at attributes. More details of the attribute masking SSL can be found in the Supplementary Materials (SM). After the GNN pre-training is finished, we then fine-tune the pre-trained GNN model on specific prediction tasks.

4. Organic Photovoltaic Dataset

We first apply the attribute masking SSL strategy to predict the quantum property of organic photovoltaic (OPV) molecules. The OPV dataset used in this work contains the SMILES strings of the 90823 unique molecules and their corresponding the ground state electronic properties obtained from DFT calculations with B3LYP/6-31G(d) (St. John et al., 2019). 5000 molecules were randomly selected from the dataset for each of the validation and test sets. The remaining data is used for pre-training and fine-tuning. The underlying GNN used is the Graph Isomorphism Network (Xu et al., 2019), a powerful GNN that is widely used in a variety of graph related task. However it is worth noting that the pre-training strategy advocated in this work is general and applicable to most GNNs. The entire training process consists of pre-training and fine-tuning. We first pre-train the GNNs with the entire training dataset (without the molecules in the test and validation sets) using the attribute masking SSL strategy as depicted in Fig. 1 (b). After pre-training, we then fine-tuned the GNNs with just a small number of the labelled data. The details of the training process and parameters can be found in the SM.

The ground state electronic properties we focus on are the values of HOMO-LUMO gap, HOMO and LUMO, the corresponding results are shown in Fig. 2 in which the test set mean absolute errors (MAEs) of the property predictions (in eV) are shown as a function of the number of labelled training data used in the fine-tuning stage. We use all 80823 unlabelled molecules for the pre-training. The prediction results with and without SSL pre-training are shown as black and blue dots with dashed lines, respectively. It can be seen that the use of SSL pre-training generally leads to considerable improvement in prediction accuracy. For example with 1000 labelled training data, the MAE for HOMO-LUMO gap prediction drops from 0.203eV to 0.144eV, an approximately 30% improvement in accuracy. From another point of view, GNN without training would need approximately four times the number of labelled training data to achieve the same performance as the GNN with pre-training. Sizable but smaller relative improvements (approximately 20% reduction in MAEs with 1000 labelled data) can also be observed for the predictions of HOMO and LUMO values.

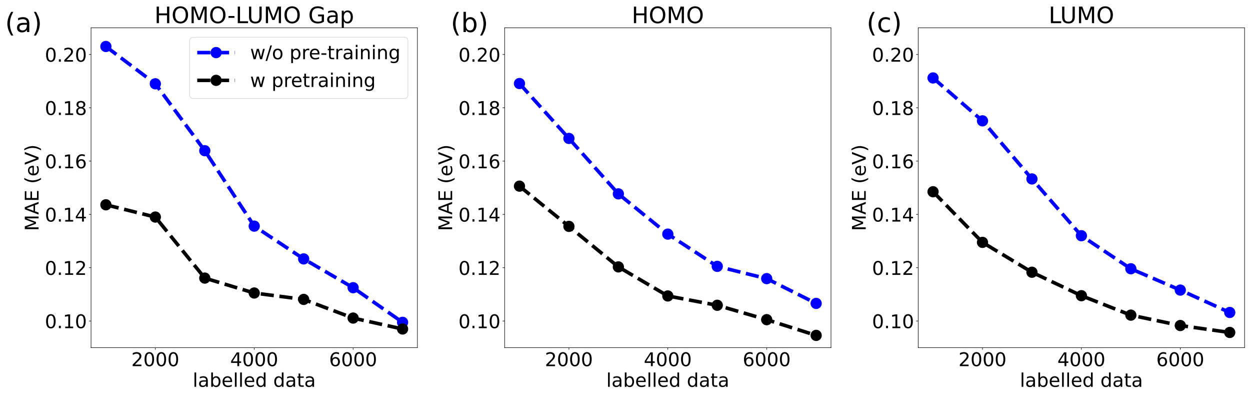

Figure 2. Test set mean absolute errors (MAE) of (a) HOMO-LUMO gap, (b) HOMO and (c) LUMO of organic photovoltaics (OPV) molecules as a function of the number of labelled training data. Blue lines represent the results from direct training of GNNs without pre-learning, whereas the black lines denote the results of GNNs pre-train with unlabelled data.

As expected, it can also be observed that the advantage from SSL pre-training decreases when the number of labelled data increases for all three properties, which shows that SSL is most useful for applications when labelled data is scarce.

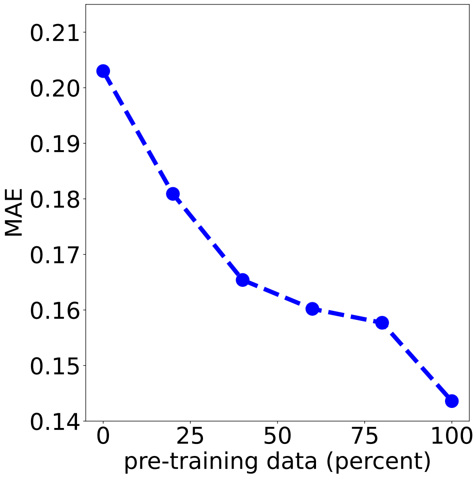

Figure 3. Test set MAE of HOMO-LUMO gap for OPV dataset as a function of the number of unlabelled pre-training data. There are a total of 80823 molecules in the pre-training dataset.

Next we investigate the performance dependence of SSL on the amount of unlabelled data used in the pre-training stage. Fig. 3 shows the test set MAE of HOMO-LUMO gap prediction as a function of the number of unlabelled pre-training data, expressed in terms of the percentage of the total pre-training dataset (80823 molecules). We use 1000 labelled data for the fine-tuning. As expected, the prediction performance is sensitive to the amount of data used in the pre-training of GNNs, more data leads to higher performance. Interestingly, it is shown that the MAE curve has not leveled even after using all of the training data for pre-training, indicating the performance of SSL can be further improved from more unlabelled data. One could potentially pre-train GNNs using other large publicly available chemical datasets such the PubChem (Kim et al., 2021) and ZINC (Irwin et al., 2012) datasets in addition to the OPV dataset, each of these datasets contains millions of molecules and could potentially further boost the performance of the GNNs. Though pre-training of GNNs with such a large dataset could be computationally intensive, it needs only to be performed once and the pre-trained GNNs can be used for various downstream tasks. Gigantic pre-trained models have been released in the computer vision (Simonyan & Zisserman, 2015; Szegedy et al., 2015; Krizhevsky et al., 2012) and nature language processing (Devlin et al., 2019; Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, 2019) domains, and had subsequently led to rapid advances in these fields, it will be a fruitful endeavor to attempt similar approach in chemical sciences in the future.

Embedding Visualization

To better understand the effect of pre-training, we use the t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten & Hinton, 2008) technique to visualize the GNN vector embeddings after SSL. t-SNE is an unsupervised dimensionality reduction technique commonly used for the visualization of high-dimensional datasets. As a non-linear dimensionality technique, t-SNE reduces the dimensions of correlated data by projecting the original set of vectors onto small number of principal components while preserving most of the data variation. Fig. 4 shows the two-dimensional t-SNE distribution of the molecule vector embeddings after performing SSL pre-training. It can be seen that after pre-training the embeddings of many molecules form clusters in the t-SNE distribution instead of being randomly distributed. It is expected that molecules in a cluster or nearby molecules in the reduced dimensions share some structural similarity. By looking into the structures of some of the adjacent molecules, we indeed find some nearby molecules that are structurally very similar, as illustrated in Fig. 4. However due to the diversity and complexity of the OPV dataset, many molecules are isolated and do not belong to any cluster in the reduced dimensions. In additional to higher prediction accuracy, the recognition of similar molecular structures during pre-training also leads to learning curves that converge faster and are more stable as compared to the learning curves without pre-training (see SM).

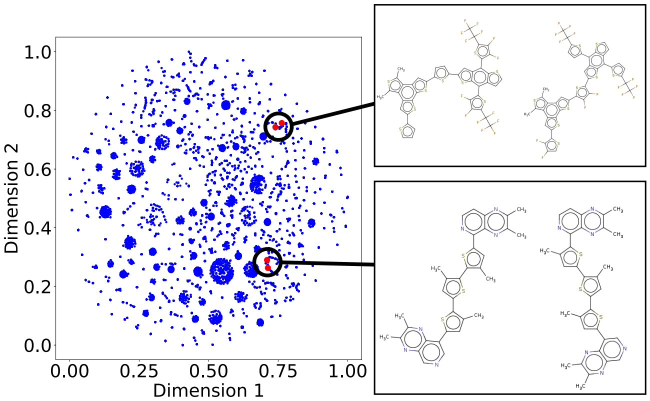

Figure 4. Two-dimensional visualization of the results of the pre-training using t-SNE.

5. Non-equilibrium configurations

We next explore the capability of SSL on molecules in non-equilibrium configurations and their excited state properties. For this purpose we use a dataset that contains 80000 non-equilibrium configurations of sexithiophene molecule. These configurations are generated from molecular dynamics simulations at 1000K and the excited properties are obtained using time-dependent DFT (TD-DFT) calculations with the CAM-B3LYP functional, more details about the dataset can be found in Refs. 45 and 40. Applications of ML for spectroscopy and exciton transport using non-equilibrium molecular configurations have previously been explored in Refs. (Lu et al., 2020; Lee et al., 2021; Farahvash et al., 2020). Similar to the OPV calculations, we randomly select 5000 configurations each for validation and testing, and use the remaining data for pre-training and fine-tuning. The underlying GNN used is SchNet since we need to take the 3D molecular coordinates into account in addition to the graph structures (Schütt et al., 2018). Since the non-equilibrium dataset involves the same molecule, the standard atom/bond type masking in SSL is not applicable. Instead, we mask the inter-atomic distance feature vectors after the rbf layer in the filter-generating network, and SchNet is tasked to predict these pair-wise vector in the pre-training stage.

In Fig. 5 we evaluate the performance of pre-training in predicting two excited state properties: the lowest excited state energy and the magnitude of its transition dipole moment. It can be seen from Fig. 5a that the prediction of the excited state energy is considerably improved by the use of SSL pre-training as compared to the results without pre-training. Similar to the results with OPV dataset, the improvement is most significant when the labelled data is scarce. For example, with only 1000 labelled data, the MAEs drops from 0.181eV to 0.142eV, an approximately 22% reduction. As the number of labelled training increases, the improvement decreases and becomes nearly negligible beyond 4000 labelled data. Next in Fig. 5b, we show the test set MAE of transition dipole moment magnitude prediction as a function of the number of labelled training data. We again observe significant improvement from SSL pre-training of GNNs. Interestingly, unlike the excited state energy prediction, the magnitude of improvement of nearly $16-20\%$ is nearly constant even when the labelled training data increases from 1000 to 5000. It has been previously shown that the prediction of transition dipole moment is more difficult compared to other electronic properties (Lu et al., 2020; Ye et al., 2019), our results here suggest SSL could be especially useful for the prediction of such challenging property in which the need for training data is greater.

Discussions and Conclusions

We demonstrate the capability of SSL pre-training of GNNs in improving the prediction accuracy of optoelectronic properties of OSCs. Similar to pre-training strategies in other domains such as computer vision where the neural networks are tasked to learn the basic features such as edges and curves in pictures from abundant unlabelled data, the SSL pre-training allows GNNs to recognize the basic structures in molecules, such as bonds and ring structures. From the t-SNE plot in Fig. 4, it can be seen that some structurally similar molecules are grouped together in the vector embedding space during the process of pre-training, such grouping assist the training process during the fine-tuning stage. Importantly, the pre-training only needs to be performed once, and the pre-trained GNNs can be fine-tuned for any property-specific task.

In this work we apply SSL to two types of problems, namely equilibrium ground state and non-equilibrium excited state property predictions. For the equilibrium case, the only input required is the SMILES strings of the molecules. A potential application is the virtual screening of OSC molecules for desirable optical properties before they are actually synthesized. For the non-equilibrium counterpart, accurate prediction of the dependence of the OSC optoelectronic

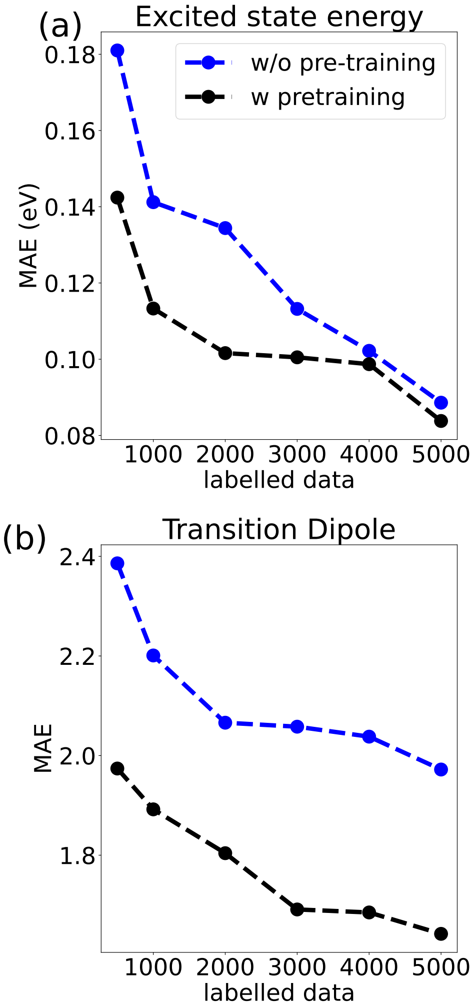

Figure 5. Evaluation on excited state properties of non-equilibrium configurations. Test set MAEs of (a) excited state energy and (b) magnitude of transition dipole moment of sexithiophene molecule as a function of the number of labelled training data. Blue lines represent the results from direct training of SchNet without pre-learning, whereas the black lines denote the results of SchNet pre-trained with unlabelled data.

properties on the 3D configuration is crucial in understanding how disorder could affect the performance of OSCs. In both cases, we obtain considerable performance improvement over results without the use of pre-training, and the improvement is most significant when labelled training is scarce. Finally, there are other ML strategies that could alleviate the need of labelled data, e.g. transfer learning and active learning, and it will be a fruitful endeavor to combine SSL with these strategies to maximize the potential of ML methods for chemical applications.

References

Source code. URL https://github.com/zaixizhang/SSL_OSC.

Adam, P., Soumith, C., Gregory, C., Edward, Y., Zachary, D., Zeming, L., Alban, D., Luca, A., and Adam, L. Automatic differentiation in pytorch. In *Proceedings of Neural Information Processing Systems*, 2017.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, I. S. Language Models are Unsupervised Multitask Learners. *Technical Report, OpenAI*, 2019. URL https://github.com/codelucas/newspaper.

Atahan-Evrenk, S. and Atalay, F. B. Prediction of Intramolecular Reorganization Energy Using Machine Learning. *The Journal of Physical Chemistry A*, 123 (36):7855–7863, sep 2019. ISSN 1089-5639. doi: 10.1021/acs.jpca.9b02733. URL https://pubs.acs.org/doi/10.1021/acs.jpca.9b02733.

Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7): 074106, feb 2011. ISSN 0021-9606. doi: 10.1063/1.3553717. URL http://aip.scitation.org/doi/10.1063/1.3553717.

Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry*, 115(16):1032–1050, aug 2015. ISSN 00207608. doi: 10.1002/qua.24890. URL http://doi.wiley.com/10.1002/qua.24890.

Behler, J. Perspective: Machine learning potentials for atomistic simulations. *The Journal of Chemical Physics*, 145(17):170901, nov 2016. ISSN 0021-9606. doi: 10.1063/1.4966192. URL http://aip.scitation.org/doi/10.1063/1.4966192.

Bian, L., Sorescu, D. C., Chen, L., White, D. L., Burkert, S. C., Khalifa, Y., Zhang, Z., Sejdic, E., and Star, A. Machine-Learning Identification of the Sensing Descriptors Relevant in Molecular Interactions with Metal Nanoparticle-Decorated Nanotube Field-Effect Transistors. *ACS Applied Materials & Interfaces*, 11(1):1219–1227, jan 2019. ISSN 1944-8244. doi: 10.1021/acsami.8b15785. URL https://pubs.acs.org/doi/10.1021/acsami.8b15785.

Chen, G., Chen, P., Hsieh, C.-Y., Lee, C.-K., Liao, B., Liao, R., Liu, W., Qiu, J., Sun, Q., Tang, J., Zemel, R., and Zhang, S. Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models. pp. arXiv:1906.09427, jun 2019. URL http://arxiv.org/abs/1906.09427arxiv:1906.09427.

Chen, W. K., Liu, X. Y., Fang, W. H., Dral, P. O., and Cui, G. Deep Learning for Nonadiabatic Excited-State Dynamics. *Journal of Physical Chemistry Letters*, 9(23):6702–6708, 2018. ISSN 19487185. doi: 10.1021/acs.jpclett.8b03026.

Christensen, A. S., Sirumalla, S. K., Qiao, Z., O'Connor, M. B., Smith, D. G. A., Ding, F., Bygrave, P. J., Anandkumar, A., Welborn, M., Manby, F. R., and Miller, T. F. OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. pp. 1–16, jul 2021a. URL http://arxiv.org/abs/2107.00299.

Christensen, A. S., Sirumalla, S. K., Qiao, Z., O'Connor, M. B., Smith, D. G. A., Ding, F., Bygrave, P. J., Anandkumar, A., Welborn, M., Manby, F. R., and Miller, T. F. OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. may 2021b. ISSN 0021-9606. doi: 10.1063/5.0061990. URL http://doi.wiley.com/10.1002/9783527685172http://arxiv.org/abs/2107.00299.

Deotare, P., Chang, W., Hontz, E., Congreve, D., Shi, L., Reusswig, P., Modtland, B., Bahlke, M., Lee, C., Willard, A. P., Bulović, V., Van Voorhis, T., and Baldo, M. Nanoscale transport of charge-transfer states in organic donor–acceptor blends. *Nature Materials*, (September):1–6, sep 2015. ISSN 1476-1122. doi: 10.1038/nmat4424. URL http://www.nature.com/doifinder/10.1038/nmat4424.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171–4186, 2019.

Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *Journal of Physical Chemistry Letters*, 11(6): 2336–2347, 2020. ISSN 19487185. doi: 10.1021/acs.jpclett.9b03664.

Dral, P. O., Barbatti, M., and Thiel, W. Nonadiabatic Excited-State Dynamics with Machine Learning. *The Journal of Physical Chemistry Letters*, 9(19):5660–5663, oct 2018. ISSN 1948-7185. doi: 10.1021/acs.jpclett.8b02469. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.8b02469.

Duan, C., Liu, F., Nandy, A., and Kulik, H. J. Semi-supervised Machine Learning Enables the Robust Detection of Multireference Character at Low Cost. *Journal of Physical Chemistry Letters*, 11(16):6640–6648, 2020. ISSN 19487185. doi: 10.1021/acs.jpclett.0c02018.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems*, pp. 2224–2232, 2015.

Farahvash, A., Lee, C.-K., Sun, Q., Shi, L., and Willard, A. P. Machine learning Frenkel Hamiltonian parameters to accelerate simulations of exciton dynamics. *The Journal of Chemical Physics*, 153(7):074111, aug 2020. ISSN 0021-9606. doi: 10.1063/5.0016009. URL http://aip.scitation.org/doi/10.1063/5.0016009.

Fey, M. and Lenssen, J. E. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1263–1272. JMLR.org, 2017.

Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T. D., Duvenaud, D., Maclaurin, D., Blood-Forsythe, M. A., Chae, H. S., Einzinger, M., Ha, D.-G. G., Wu, T., Markopoulos, G., Jeon, S., Kang, H., Miyazaki, H., Numata, M., Kim, S., Huang, W., Hong, S. I., Baldo, M., Adams, R. P., and Aspuru-Guzik, A. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials*, 15(10):1120–1127, oct 2016. ISSN 14764660. doi: 10.1038/nmat4717. URL http://www.nature.com/articles/nmat4717.

Hachmann, J., Olivares-Amaya, R., Atahan-Evrenk, S., Amador-Bedolla, C., Sánchez-Carrera, R. S., Gold-Parker, A., Vogt, L., Brockway, A. M., and Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *The Journal of Physical Chemistry Letters*, 2(17):2241–2251, sep 2011. ISSN 1948-7185. doi: 10.1021/jz200866s. URL https://pubs.acs.org/doi/10.1021/jz200866s.

Hains, A. W., Liang, Z., Woodhouse, M. A., and Gregg, B. A. Molecular Semiconductors in Organic Photovoltaic Cells. *Chemical Reviews*, 110(11):6689–6735, nov 2010. ISSN 0009-2665. doi: 10.1021/cr9002984. URL http://pubs.acs.org/doi/abs/10.1021/cr9002984.

Hao, Z., Lu, C., Huang, Z., Wang, H., Hu, Z., Liu, Q., Chen, E., and Lee, C. ASGN: An Active Semi-supervised Graph Neural Network for Molecular Property Prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data*

Mining, pp. 731–752, New York, NY, USA, aug 2020. ACM. ISBN 9781450379984. doi: 10.1145/3394486.3403117. URL https://dl.acm.org/doi/10.1145/3394486.3403117.

Hedley, G. J., Ruseckas, A., and Samuel, I. D. W. Light Harvesting for Organic Photovoltaics. *Chemical Reviews*, 117 (2):796–837, jan 2017. ISSN 0009-2665. doi: 10.1021/acs.chemrev.6b00215. URL https://pubs.acs.org/doi/10.1021/acs.chemrev.6b00215.

Hu*, W., Liu*, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for Pre-training Graph Neural Networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, 2012. ISSN 15499596. doi: 10.1021/ci3001277.

Janai, M. A. B., Woon, K. L., and Chan, C. S. Design of efficient blue phosphorescent bottom emitting light emitting diodes by machine learning approach. *Organic Electronics*, 63(July):257–266, dec 2018. ISSN 15661199. doi: 10.1016/j.orgel.2018.09.029. URL https://doi.org/10.1016/j.orgel.2018.09.029https://linkinghub.elsevier.com/retrieve/pii/S1566119918304944.

Jørgensen, P. B., Mesta, M., Shil, S., García Lastra, J. M., Jacobsen, K. W., Thygesen, K. S., and Schmidt, M. N. Machine learning-based screening of complex molecules for polymer solar cells. *The Journal of Chemical Physics*, 148(24):241735, jun 2018. ISSN 0021-9606. doi: 10.1063/1.5023563. URL http://dx.doi.org/10.1063/1.5023563http://aip.scitation.org/doi/10.1063/1.5023563.

Kanal, I. Y., Owens, S. G., Bechtel, J. S., and Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *The Journal of Physical Chemistry Letters*, 4(10):1613–1623, may 2013. ISSN 1948-7185. doi: 10.1021/jz400215j. URL https://pubs.acs.org/doi/10.1021/jz400215j.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, jan 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa971. URL https://academic.oup.com/nar/article/49/D1/D1388/5957164.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2015.

Klicpera, J., Groß, J., and Günnemann, S. Directional Message Passing for Molecular Graphs. In *International Conference on Learning Representations*, mar 2020. URL http://arxiv.org/abs/2003.03123.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kulichenko, M., Smith, J. S., Nebgen, B., Li, Y. W., Fedik, N., Boldyrev, A. I., Lubbers, N., Barros, K., and Tretiak, S. The Rise of Neural Networks for Materials and Chemical Dynamics. *The Journal of Physical Chemistry Letters*, 12(26):6227–6243, jul 2021. ISSN 1948-7185. doi: 10.1021/acs.jpclett.1c01357. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.1c01357.

Landrum, G. and Others. RDKit: Open-source cheminformatics. 2006.

Lederer, J., Kaiser, W., Mattoni, A., and Gagliardi, A. Machine Learning–Based Charge Transport Computation for Pentacene. *Advanced Theory and Simulations*, 2(2):1800136, feb 2019. ISSN 2513-0390. doi: 10.1002/adts.201800136. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adts.201800136.

Lee, C. K., Shi, L., and Willard, A. P. Modeling the Influence of Correlated Molecular Disorder on the Dynamics of Excitons in Organic Molecular Semiconductors. *The Journal of Physical Chemistry C*, 123(1):306–314, jan 2019. ISSN 1932-7447. doi: 10.1021/acs.jpcc.8b11504. URL https://pubs.acs.org/doi/10.1021/acs.jpcc.8b11504.

Lee, C. K., Lu, C., Yu, Y., Sun, Q., Hsieh, C. Y., Zhang, S., Liu, Q., and Shi, L. Transfer learning with graph neural networks for optoelectronic properties of conjugated oligomers. *Journal of Chemical Physics*, 154(2), 2021. ISSN 10897690. doi: 10.1063/5.0037863. URL https://doi.org/10.1063/5.0037863.

Lee, M. Insights from Machine Learning Techniques for Predicting the Efficiency of Fullerene Derivatives-Based Ternary Organic Solar Cells at Ternary Blend Design. *Advanced Energy Materials*, 9(26):1900891, jun 2019.

ISSN 1614-6832. doi: 10.1002/aenm.201900891. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aenm.201900891.

Li, H., Zhong, Z., Li, L., Gao, R., Cui, J., Gao, T., Hu, L. H., Lu, Y., Su, Z.-M., and Li, H. A cascaded QSAR model for efficient prediction of overall power conversion efficiency of all-organic dye-sensitized solar cells. *Journal of Computational Chemistry*, 36(14):1036–1046, may 2015. ISSN 01928651. doi: 10.1002/jcc.23886. URL http://doi.wiley.com/10.1002/jcc.23886.

Liu, Z., Lin, L., Jia, Q., Cheng, Z., Jiang, Y., Guo, Y., and Ma, J. Transferable Multi-level Attention Neural Network for Accurate Prediction of Quantum Chemistry Properties via Multi-task Learning. pp. chemRxiv:12588170, 2020. doi: 10.26434/chemrxiv.12588170.v1. URL https://chemrxiv.org/articles/preprint/Transferable_Multi-level_Attention_Neural_Network_for_Accurate_Prediction_of_Quantum_Chemistry_Properties_via_Multi-task_Learning/12588170.

Lu, C., Liu, Q., Wang, C., Huang, Z., Lin, P., and He, L. Molecular Property Prediction: A Multilevel Quantum Interactions Modeling Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:1052–1060, jul 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33011052. URL https://aaai.org/ojs/index.php/AAAI/article/view/3896.

Lu, C., Liu, Q., Sun, Q., Hsieh, C.-Y., Zhang, S., Shi, L., and Lee, C.-K. Deep Learning for Optoelectronic Properties of Organic Semiconductors. *The Journal of Physical Chemistry C*, 124(13):7048–7060, apr 2020. ISSN 1932-7447. doi: 10.1021/acs.jpcc.0c00329. URL https://pubs.acs.org/doi/10.1021/acs.jpcc.0c00329.

Lu, L., Zheng, T., Wu, Q., Schneider, A. M., Zhao, D., and Yu, L. Recent Advances in Bulk Heterojunction Polymer Solar Cells. *Chemical Reviews*, 115(23):12666–12731, dec 2015. ISSN 0009-2665. doi: 10.1021/acs.chemrev.5b00098. URL https://pubs.acs.org/doi/10.1021/acs.chemrev.5b00098.

Lu, Y., Jiang, X., Fang, Y., and Shi, C. Learning to Pretrain Graph Neural Networks. *AAAI*, (2), 2021. URL https://github.com/rootlu/L2P-GNN.

Mahapatra, N., Ben-Cohen, A., Vaknin, Y., Henning, A., Hayon, J., Shimanovich, K., Greenspan, H., and Rosenwaks, Y. Electrostatic Selectivity of Volatile Organic Compounds Using Electrostatically Formed Nanowire Sensor. *ACS Sensors*, 3(3):709–715, mar 2018. ISSN 2379-3694. doi: 10.1021/acssensors.8b00044. URL https://pubs.acs.org/doi/10.1021/acssensors.8b00044.

Minaev, B., Baryshnikov, G., and Agren, H. Principles of phosphorescent organic light emitting devices. *Phys. Chem. Chem. Phys.*, 16(5):1719–1758, 2014. ISSN 1463-9076. doi: 10.1039/C3CP53806K. URL http://xlink.rsc.org/?DOI=C3CP53806K.

Musil, F., De, S., Yang, J., Campbell, J. E., Day, G. M., and Ceriotti, M. Machine learning for the structure–energy–property landscapes of molecular crystals. *Chemical Science*, 9(5):1289–1300, 2018. ISSN 2041-6520. doi: 10.1039/C7SC04665K. URL http://xlink.rsc.org/?DOI=C7SC04665K.

Myers, J. D. and Xue, J. Organic Semiconductors and their Applications in Photovoltaic Devices. *Polymer Reviews*, 52(1):1–37, jan 2012. ISSN 1558-3724. doi: 10.1080/15583724.2011.644368. URL http://www.tandfonline.com/doi/abs/10.1080/15583724.2011.644368.

Nagasawa, S., Al-Naamani, E., and Saeki, A. Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. *The Journal of Physical Chemistry Letters*, 9(10):2639–2646, may 2018. ISSN 1948-7185. doi: 10.1021/acs.jpclett.8b00635. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.8b00635.

Noé, F., Tkatchenko, A., Müller, K.-R., and Clementi, C. Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry*, 71 (1):361–390, apr 2020. ISSN 0066-426X. doi: 10.1146/annurev-physchem-042018-052331. URL https://www.annualreviews.org/doi/10.1146/annurev-physchem-042018-052331.

Olivares-Amaya, R., Amador-Bedolla, C., Hachmann, J., Atahan-Evrenk, S., Sánchez-Carrera, R. S., Vogt, L., and Aspuru-Guzik, A. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy & Environmental Science*, 4(12):4849, 2011. ISSN 1754-5692. doi: 10.1039/c1ee02056k. URL http://xlink.rsc.org/?DOI=c1ee02056k.

Ostroverkhova, O. Organic Optoelectronic Materials: Mechanisms and Applications. *Chemical Reviews*, 116(22):13279–13412, nov 2016. ISSN 0009-2665. doi: 10.1021/acs.chemrev.6b00127. URL http://pubs.acs.org/doi/abs/10.1021/acs.chemrev.6b00127.

Padula, D. and Troisi, A. Concurrent Optimization of Organic Donor-Acceptor Pairs through Machine Learning.

Advanced Energy Materials, 9(40):1902463, oct 2019. ISSN 1614-6832. doi: 10.1002/aenm.201902463. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aenm.201902463.

Padula, D., Simpson, J. D., and Troisi, A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Materials Horizons*, 6(2):343–349, 2019. ISSN 20516355. doi: 10.1039/c8mh01135d. URL http://xlink.rsc.org/?DOI=C8MH01135D.

Pereira, F., Xiao, K., Latino, D. A. R. S., Wu, C., Zhang, Q., and Aires-de Sousa, J. Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *Journal of Chemical Information and Modeling*, 57(1):11–21, jan 2017. ISSN 1549-9596. doi: 10.1021/acs.jcim.6b00340. URL https://pubs.acs.org/doi/10.1021/acs.jcim.6b00340.

Poltavsky, I. and Tkatchenko, A. Machine Learning Force Fields: Recent Advances and Remaining Challenges. *The Journal of Physical Chemistry Letters*, 12(28):6551–6564, jul 2021. ISSN 1948-7185. doi: 10.1021/acs.jpclett.1c01204. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.1c01204.

Prezhdo, O. V. Advancing Physical Chemistry with Machine Learning. *The Journal of Physical Chemistry Letters*, 11 (22):9656–9658, nov 2020. ISSN 1948-7185. doi: 10.1021/acs.jpclett.0c03130. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.0c03130.

Pyzer-Knapp, E. O., Simm, G. N., and Aspuru Guzik, A. A Bayesian approach to calibrating high-throughput virtual screening results and application to organic photovoltaic materials. *Materials Horizons*, 3(3):226–233, 2016. ISSN 2051-6347. doi: 10.1039/C5MH00282F. URL http://xlink.rsc.org/?DOI=C5MH00282F.

Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R., and Miller, T. F. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of Chemical Physics*, 153(12):124111, sep 2020. ISSN 0021-9606. doi: 10.1063/5.0021955. URL http://arxiv.org/abs/2007.08026http://aip.scitation.org/doi/10.1063/5.0021955.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, dec 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL http://www.nature.com/articles/sdata201422.

Roch, L. M., Saikin, S. K., Häse, F., Friederich, P., Goldsmith, R. H., León, S., and Aspuru-Guzik, A. From Absorption Spectra to Charge Transfer in Nanoaggregates of Oligomers with Machine Learning. *ACS Nano*, 14(6): 6589–6598, jun 2020. ISSN 1936-0851. doi: 10.1021/acsnano.0c00384. URL https://pubs.acs.org/doi/10.1021/acsnano.0c00384.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 2020-Decem(NeurIPS):1–18, 2020. ISSN 10495258.

Sahu, H., Rao, W., Troisi, A., and Ma, H. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Advanced Energy Materials*, 8(24):1801032, aug 2018. ISSN 16146832. doi: 10.1002/aenm.201801032. URL http://doi.wiley.com/10.1002/aenm.201801032.

Sajeev, R., Athira, R. S., Nufail, M., Jinu Raj, K. R., Rakhila, M., Nair, S. M., Abdul Jaleel, U. C., and Manuel, A. T. Computational predictive models for organic semiconductors. *Journal of Computational Electronics*, 12(4):790–795, dec 2013. ISSN 1569-8025. doi: 10.1007/s10825-013-0486-3. URL http://link.springer.com/10.1007/s10825-013-0486-3.

Schütt, K., Kindermans, P.-J., Felix, H. E. S., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, pp. 991–1001, 2017a.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., and Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8 (1):13890, apr 2017b. ISSN 2041-1723. doi: 10.1038/ncomms13890. URL http://www.nature.com/articles/ncomms13890.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. SchNet - A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24): 241722, jun 2018. ISSN 0021-9606. doi: 10.1063/1.5019779. URL http://aip.scitation.org/doi/10.1063/1.5019779.

Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K.-R., and Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1):5024, dec 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12875-2. URL http://arxiv.

org/abs/1906.10033http://www.nature.com/articles/s41467-019-12875-2.

Shi, L., Lee, C. K., and Willard, A. P. The Enhancement of Interfacial Exciton Dissociation by Energetic Disorder Is a Nonequilibrium Effect. *ACS Central Science*, 3 (12):1262–1270, dec 2017. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00404. URL https://pubs.acs.org/doi/10.1021/acscentsci.7b00404.

Shu, Y. and Levine, B. G. Simulated evolution of fluorophores for light emitting diodes. *The Journal of Chemical Physics*, 142(10):104104, mar 2015. ISSN 0021-9606. doi: 10.1063/1.4914294. URL http://dx.doi.org/10.1063/1.4914294http://aip.scitation.org/doi/10.1063/1.4914294.

Simine, L., Allen, T. C., and Rossky, P. J. Predicting optical spectra for optoelectronic polymers using coarse-grained models and recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(25):13945–13948, jun 2020. ISSN 0027-8424. doi: 10.1073/pnas.1918696117. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1918696117.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–14, 2015.

Sirringhaus, H. 25th Anniversary Article: Organic Field-Effect Transistors: The Path Beyond Amorphous Silicon. *Advanced Materials*, 26(9):1319–1335, mar 2014. ISSN 09359648. doi: 10.1002/adma.201304346. URL http://doi.wiley.com/10.1002/adma.201304346.

Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A. E. Less is more: Sampling chemical space with active learning. *Journal of Chemical Physics*, 148(24), 2018. ISSN 00219606. doi: 10.1063/1.5023802.

St. John, P. C., Phillips, C., Kemper, T. W., Wilson, A. N., Guan, Y., Crowley, M. F., Nimlos, M. R., and Larsen, R. E. Message-passing neural networks for high-throughput polymer screening. *The Journal of Chemical Physics*, 150(23):234111, jun 2019. ISSN 0021-9606. doi: 10.1063/1.5099132. URL http://arxiv.org/abs/1807.10363http://aip.scitation.org/doi/10.1063/1.5099132.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9. IEEE, jun 2015. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7298594. URL http://ieeexplore.ieee.org/document/7298594/.

Taylor, M. G., Nandy, A., Lu, C. C., and Kulik, H. J. Deciphering Cryptic Behavior in Bimetallic Transition-Metal Complexes with Machine Learning. *The Journal of Physical Chemistry Letters*, 12(40):9812–9820, oct 2021. ISSN 1948-7185. doi: 10.1021/acs.jpclett.1c02852. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.1c02852.

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. *arXiv preprint arXiv:2008.03226*, 2020.

Unke, O. T. and Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *Journal of Chemical Theory and Computation*, 15(6):3678–3693, jun 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.9b00181. URL https://pubs.acs.org/doi/10.1021/acs.jctc.9b00181.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

von Lilienfeld, O. A., Müller, K.-R., and Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry*, 4(7):347–358, jul 2020. ISSN 2397-3358. doi: 10.1038/s41570-020-0189-9. URL http://www.nature.com/articles/s41570-020-0189-9.

Wang, B., Chu, W., Tkatchenko, A., and Prezhdo, O. V. Interpolating Nonadiabatic Molecular Dynamics Hamiltonian with Artificial Neural Networks. *The Journal of Physical Chemistry Letters*, 12(26):6070–6077, jul 2021. ISSN 1948-7185. doi: 10.1021/acs.jpclett.1c01645. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.1c01645.

Wang, L., Huang, Y., Hou, Y., Zhang, S., and Shan, J. Graph attention convolution for point cloud semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:10288–10297, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01054.

Welborn, M., Cheng, L., and Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *Journal of Chemical Theory and Computation*, 14(9):4772–4779, sep 2018. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b00636. URL https://pubs.acs.org/doi/10.1021/acs.jctc.8b00636.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. ISSN 2041-6520. doi: 10.1039/C7SC02664A. URL http://xlink.rsc.org/?DOI=C7SC02664A.

Xie, Y., Xu, Z., Zhang, J., Wang, Z., and Ji, S. Self-Supervised Learning of Graph Neural Networks: A Unified Review. pp. 1–18, 2021. URL http://arxiv.org/abs/2102.10757.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How Powerful are Graph Neural Networks? In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Xu, R.-P., Li, Y.-Q., and Tang, J.-X. Recent advances in flexible organic light-emitting diodes. *Journal of Materials Chemistry C*, 4(39):9116–9142, 2016. ISSN 2050-7526. doi: 10.1039/C6TC03230C. URL http://xlink.rsc.org/?DOI=C6TC03230C.

Ye, S., Hu, W., Li, X., Zhang, J., Zhong, K., Zhang, G., Luo, Y., Mukamel, S., and Jiang, J. A neural network protocol for electronic excitations of N -methylacetamide. *Proceedings of the National Academy of Sciences*, 116(24): 201821044, may 2019. ISSN 0027-8424. doi: 10.1073/pnas.1821044116. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1821044116.

Zhang, Z., Liu, Q., Wang, H., Lu, C., and Lee, C.-k. Motif-based Graph Self-Supervised Learning for Molecular Property Prediction. (NeurIPS):1–13, 2021.

A. Implementation Details of Graph Neural Networks

A.1. Implementation of GIN

We apply Graph Isomorphism Network (GIN) (Xu et al., 2019) for the molecular property prediction of OPV dataset. We select the following settings for GIN: 300 dimensional hidden units, 5 GNN layers, ReLU activation, dropout probability 0.5 for all layers except the input layer, and average pooling for the readout function.

For the input to GIN, we use RDKit (Landrum & Others, 2006) to obtain the node and edge features: Node features: Atomic number; Chirality tag: {unspecified, tetrahedral cw, tetrahedral ccw, other}. Edge features: Bond type: {single, double, triple, aromatic}; Bond direction: {–, endupright, enddownright}.

GIN is trained with Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.001. Both pretraining and finetuning are trained for 100 epochs. For self-supervised pretraining, we use a batch size of 256, while for finetuning, we use a batch size of 32. We use Pytorch (Adam et al., 2017) and Pytorch Geometric (Fey & Lenssen, 2019) for all of our implementation.

A.2. Implementation of SchNet

We apply SchNet (Schütt et al., 2017a) for the property prediction of non-equilibrium thiophene molecules. We select the following settings for SchNet: 2 interaction blocks, 64 dimensional hidden units, shifted softplus activation, 3.0 cutoff, 0.1 width, and sum pooling for the readout function. The input to SchNet are the atom attributes, i.e. atomic numbers and the interatomic distances.

SchNet is trained with Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0001. We pretrain SchNet for 100 epochs and finetune it for 1000 epochs. For both self-supervised pre-training and finetuning, we use a batch size of 20. We use PyTorch (Adam et al., 2017) and PyTorch Geometric (Fey & Lenssen, 2019) for all of our implementation.

For both GIN and SchNet, we use mean squared error (MSE) as loss function in the training process. We use the same learning rates as the original papers, other hyperparameters are found using grid search. The GNNs are implemented in PyTorch 1.8.0.

B. Implementation Details of Self-supervised Learning

We employ attribute masking as the strategy for self-supervised learning (Hu* et al., 2020). Generally, we mask the node/edge attributes by randomly initializing the embed-

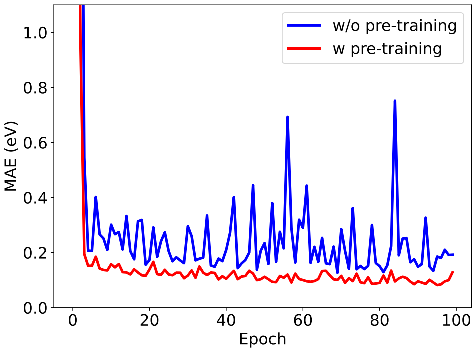

Figure 6. Training curves of HOMO with (red) and without (blue) SSL pre-training. A total of 80823 molecules are used for pre-training, and 1000 labeled data are used for fine-tuning.

dings, and then let GNNs predict those attributes based on neighboring structure. In experiments, we randomly sample 15% nodes and edges and replace them with special mask indicators. For SchNet, we replace the interatomic distance embeddings of m asked edges after the rbf layer with special mask indicators. We then apply GNNs to obtain the corresponding node/edge embeddings (edge embeddings can be obtained as a sum of node embeddings of the edge's end nodes). Finally, a fully-connected layer is applied on top of embeddings to predict the masked node/edge attributes. For the masked edges in SchNet, we predict which interval the interatomic distance belongs to. The pre-training step takes about 4 hours whereas the the fine-tuning step takes about 1 hour, both on a single NVIDIA V100 GPU.

C. Dependence on Pre-training Data Size

In Fig.7, we show additional results of the performance dependence of SSL on the amount of unlabelled data used in the pre-training stage. Similar to Fig. 3 in the main text, 1000 labelled data is used for the fine-tuning stage.

D. Comparing GNN predicted values with DFT labels

In Fig.8, we compare the GNN predicted values of HOMO against the values from DFT calculations for the OPV dataset. With pre-training, the R^2 coefficient increases from 0.856 to 0.928. It is also observed that the improvement is most noticeable when the absolute HOMO values are large.

E. Learning Curves

In Fig. 6, we show the training set MAEs of HOMO as a function of epoch number for the OPV dataset. It can be seen that the use of SSL pre-training not only leads to lower MAE, the training curve is also less noisy.

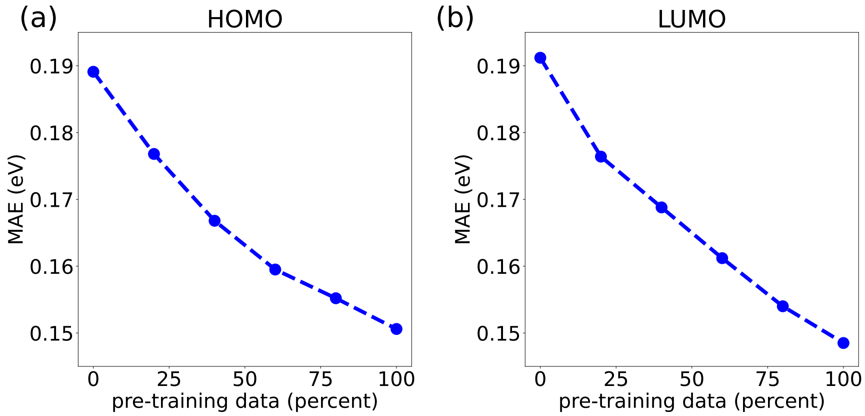

Figure 7. Test set MAEs of HOMO and LUMO for OPV dataset as a function of the number of unlabelled pre-training data. There are a total of 80823 molecules in the pre-training dataset.

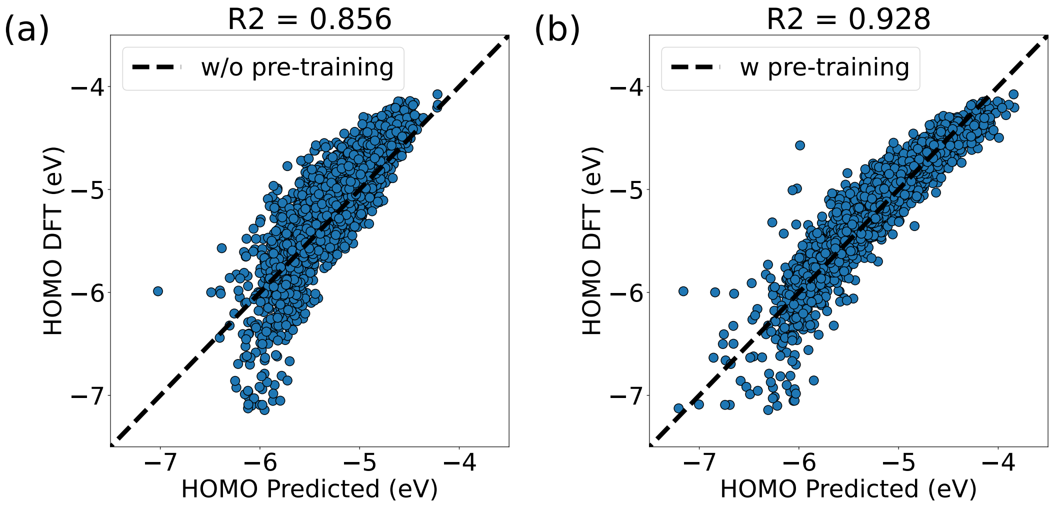

Figure 8. Scatter plots comparing the GNN predicted values of HOMO against DFT calculations without (a) and with pre-training (b). A total of 80823 molecules are used for pre-training, and 1000 labeled data are used for fine-tuning