Accelerate the screening of complex materials by learning to reduce random and systematic errors

Tian Xie, Arthur France-Lanord, Yanming Wang, Jeffrey Lopez, Michael Austin Stolberg, Megan Hill, Graham Michael Leverick, Rafael Gomez-Bombarelli, Jeremiah A. Johnson, Yang Shao-Horn, Jeffrey C. Grossman
Massachusetts Institute of Technology
Cambridge, MA 02139
txie@mit.edu, jcg@mit.edu

Abstract

Graph neural networks have been widely adopted to accelerate the screening of materials. Most existing studies assume that the ground truth data are generated through a deterministic, unbiased process, which might break down for the simulation of complex materials. In this work, we find that a multi-task graph neural network that learns from a large number of biased, noisy data and a small number of unbiased data can reduce both random and systematic errors. This allows us to use cheaper, unconverged simulations to accelerate the screening of a class of polymer electrolyte materials by 22.8 times.

1 Introduction

Graph neural networks have been widely adopted for accelerating the discovery of molecular [1–5] and solid materials [6–8] under the supervised learning framework. Most of these studies assume that the ground truth data are generated through a deterministic, unbiased process, such as performing quantum mechanical simulations[1]. However, this assumption might break down for more complex materials and properties. First, the atomic structure of the material, i.e. inputs to a simulator, can only be sampled via a stochastic process for complex materials like polymers and solids with defects. Second, some property simulators, like molecular dynamics (MD), are intrinsically stochastic and have large uncertainties on simulated properties. Finally, with a limited computation budget, the simulation of complex materials may not converge which leads to systematic errors.

In this work, we aim to study how graph neural networks perform on predicting material properties when the training data have significant random and systematic errors. In practice, researchers usually simulate complex materials with multiple independent simulations on same material to reduce random errors, and long-time simulation to reduce systematic errors. [9–11] We hope to demonstrate that these errors can be reduced by learning from a large number of biased, noisy data and a small number of unbiased data, reducing the need of redundant simulations.

We focus on a specific material design problem of discovering polymer electrolytes with a higher ionic conductivity for lithium ion batteries[12–14]. Due to the computational cost to reduce random and systematic errors, the previous computational studies for this problem only simulated around 10 materials [9–11]. In the scope of this class of materials, we hope to answer the following questions:

- By learning from data with large random errors, can a graph neural network predict the *true* property with errors smaller than the random errors from data?

[1] In principle, quantum mechanical simulations like density functional theory are also stochastic. But the uncertainty is small and usually neglected. They are also biased estimation of the true experimental property, which is often not discussed in the context of supervised learning.

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

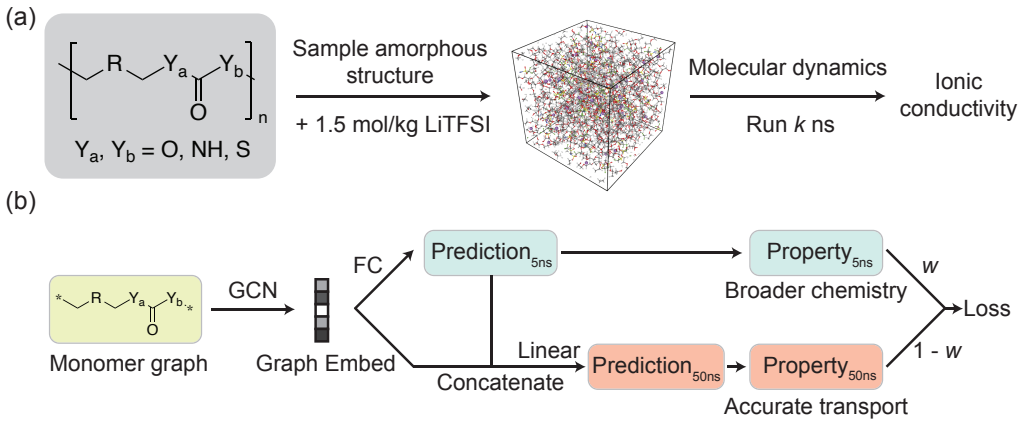

Figure 1: (a) Illustration of the data generalization process. Monomers are sampled from a pharmaceutical database [15] to ensure synthesizability. (b) Multi-task learning architecture to reduce the random and systematic errors from simulations.

- By learning from a large number of biased data and a small number of unbiased data, can a multi-task graph neural network learn to reduce systematic errors from biased data?
- Combining the reduction of random and systematic errors, how much acceleration can we achieve for screening polymer electrolytes?

We find affirmative answers for the first two questions. Using the combined model that reduces the random and systematic errors, we successfully screened a polymer space including 6247 materials, significantly larger than previous works. We believe that the ability of graph neural networks to reduce random and systematic errors have broad implications for the screening of complex materials, because the simulation of these materials often suffer from the large computational cost similar to the polymer electrolytes.

2 Methods

2.1 Data generation

Polymer datasets As shown in Fig. 1(a), the ionic conductivity of the polymers are simulated via a two-step process. The amorphous structure of the polymer is first sampled with a Monte Carlo algorithm, and then the equilibrium structure is simulated with MD for k ns. We generate two datasets by running two types of simulations: 1) $5\,\mathrm{ns}$ dataset: we sample equilibrium structure and run $5\,\mathrm{ns}$ MD, obtaining conductivity for 876 polymers; 2) $50\,\mathrm{ns}$ dataset: we use the equilibrium structure from $5\,\mathrm{ns}$ dataset and run a much more expensive $50\,\mathrm{ns}$ MD, obtaining conductivity for 117 polymers.

For both datasets, there are significant random errors for simulated properties. Between these two datasets, the $5\,\mathrm{ns}$ properties have systematic errors with respect to the $50\,\mathrm{ns}$ properties since the former is not converged, but the random errors between them are highly correlated because they begin with the same initial configuration.

Toy datasets We do not have access to the *true* property in the polymer datasets, which would in principle require averaging multiple $50\,\mathrm{ns}$ simulations. To evaluate the model performance with respect the true properties, we use the same polymers from $5\,\mathrm{ns}$ dataset and compute the LogP of each polymer using Crippen's approach [16, 17], which is a deterministic simulator. Then, we add different levels of gaussian random noises into the LogP values to imitate the random errors in simulated conductivities.

Table 1: Random error reduction in the toy datasets with different noises and the $5\,\mathrm{ns}$ dataset

Dataset	True RMSE	Apparent RMSE	Noise Std	Estimated true RMSE
Toy (Std = 0)	0.048	0.048	0	0.048
Toy (Std = 0.08)	0.071	0.105	0.08	0.068
Toy (Std = 0.32)	0.128	0.334	0.32	0.096
Toy (Std = 1.28)	0.346	1.20	1.28	NaN
Toy (Std = 5.12)	0.517	5.18	5.12	0.767
$5\,\mathrm{ns}$ polymer	-	0.145	0.117	0.085

2.2 Network architecture

We employ a multi-task graph neural network to reduce both random and systematic errors as shown in Fig. 1(b). We first encode the monomer structure as a graph $\mathcal{G}$ and use a graph neural network G [6] to learn a representation for the corresponding polymer, $\boldsymbol{v}_\mathcal{G} = G(\mathcal{G})$.

To reduce random errors, we use the robustness of neural networks against random noises in the training data, previously demonstrated in images [18] and graphs [19]. We assume that the computed target property (e.g. conductivity) has a shared random error ϵ over the true property $f(\mathcal{G})$,

$$t = f(\mathcal{G}) + \epsilon, \tag{1}$$

where f is a deterministic function mapping from molecular graph to true property, and ϵ is a random variable independent of $\mathcal{G}$ with zero mean. Note that ϵ should be a function of $\mathcal{G}$ in principle, but similar noises in observed across polymers. By regressing over t, it is possible to learn $f(\mathcal{G})$ even when the noises are large [18] if we have enough training data. It is only possible to generate enough training data with $5\,\mathrm{ns}$ simulations, so we use a feed forward network g_1 to predict $t_{5\,\mathrm{ns}}$ with the graph representation,

$$y_{5\,\mathrm{ns}} = g_1(\boldsymbol{v}_\mathcal{G}), \tag{2}$$

which aims to learn an approximation to the true property function $f_{5\,\mathrm{ns}}$ despite the random errors. However, there is a systematic error between $f_{5\,\mathrm{ns}}$ and $f_{50\,\mathrm{ns}}$ due to the slow relaxation of polymers. To correct this error, we perform a small amount of $50\,\mathrm{ns}$ simulation to generate data for the converged conductivities. This correction can then be learned with a linear layer g_2 using both predictions from $5\,\mathrm{ns}$ simulations and the graph representations,

$$y_{50\,\mathrm{ns}} = g_2(\boldsymbol{v}_\mathcal{G} \parallel y_{5\,\mathrm{ns}}), \tag{3}$$

where $\parallel$ denotes concatenation.

Finally, the larger $5\,\mathrm{ns}$ dataset and the smaller $50\,\mathrm{ns}$ dataset can be trained jointly using a combined loss function,

$$\mathrm{Loss} = (1-w)\cdot\frac{1}{N_{5\,\mathrm{ns}}}\sum_{\mathcal{G}_{5\,\mathrm{ns}}}(y_{5\,\mathrm{ns}} - t_{5\,\mathrm{ns}})^2 + w\cdot\frac{1}{N_{50\,\mathrm{ns}}}\sum_{\mathcal{G}_{50\,\mathrm{ns}}}(y_{50\,\mathrm{ns}} - t_{50\,\mathrm{ns}})^2, \tag{4}$$

where w is a weight between 0 and 1.

3 Experiments

3.1 Random error reduction

We first study the toy dataset which we have access to the *true* property $f(\mathcal{G})$ in Eq. 1. We train the $5\,\mathrm{ns}$ branch of our network, i.e. set $w = 0$, with the toy dataset with different levels of noises. Table 1 shows the true root mean squared errors (RMSEs) with respect to the original LogPs and apparent RMSEs with respect to the noisy LogPs, using the toy dataset with different noise levels. We observe that the true RMSEs become smaller than the noise standard deviation when it is larger than 0.08. This result shows that, on average, our model predicts LogP more accurately than performing a simulation of LogP due to the existence of large noises in the simulation.

However, we do not have access to the *true* property for the $5\,\mathrm{ns}$ polymer dataset. So we can only compute the apparent RMSE, not the true RMSE. To estimate the true RMSE, we assume that the

Table 2: Systematic error reduction between different methods.

Method	MAE
No correction	0.528
Linear	0.152
Single-task	0.137 ± 0.025
Multi-task	0.093 ± 0.017

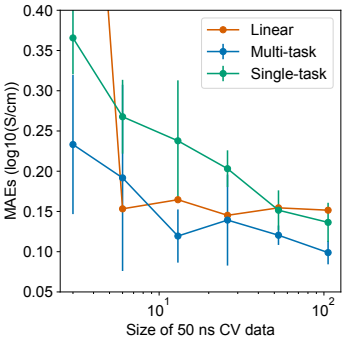

Figure 1: Performances of the methods with less training data from 50 ns dataset.

random errors for 5 ns MD conductivity follows a gaussian distribution. We can estimate the true mean squared error (MSE) with

$$\mathrm{MSE}(y, f(\mathcal{G})) = \mathrm{MSE}(y, t) - \mathop{\mathbb{E}}_{\mathcal{G}}[\epsilon^2], \tag{5}$$

where $\mathrm{MSE}(y, t)$ is the apparent MSE and $\mathbb{E}_{\mathcal{G}}[\epsilon^2]$ is the variance of the gaussian noise (detailed derivation in the appendice A). For the toy datasets, we find that this estimate gives results that are close to the true RMSE, although some differences exist due to the relative small size of the datasets. Our estimated true RMSE for the 5 ns polymer dataset is $0.085\,\log_{10}(\mathrm{S/cm})$, smaller than the standard deviation of random noise for running 5 ns simulations $0.117\,\log_{10}(\mathrm{S/cm})$.

3.2 Systematic error reduction

To learn the systematic differences between 5 ns and 50 ns datasets, we co-train our model with both datasets using 10-fold cross validations and $w = 10^{-4}$ to predict 50 ns properties. We compare the performance of our model with several baselines: 1) Linear. This baseline learns a linear model to predict 50 ns properties from 5 ns properties, which does not consider the differences between polymers; 2) Single-task. This baseline only uses the 50 ns branch of the model ($w = 1$), which directly predicts properties from molecular structure instead of learning errors.

In Table 2, we find that our multi-task model outperforms both baselines. This shows that our model learns a customized correction to each polymer, which performs better than an overall linear correction to all polymers. In Fig. 1, we explore the performance of our model with less training data from the 50 ns dataset. Although with large uncertainty, the performance of the multi-task model decreases relatively slowly with less training data, and it seems to still have some correction ability even with 13 training data points. This shows the advantage of co-training a larger 5 ns dataset and a smaller 50 ns dataset – it is much easier to learn a systematic correction than predicting the property from scratch. In contract, the performance of a single-task model directly predicting 50 ns conductivity degrades much faster with less training data.

4 Estimation of the acceleration

The ability for a multi-task graph neural network to reduce random and systematic errors indicates that redundant calculations that are performed on individual complex materials may not be necessary in the context of material screening. To screen a space of 6247 polymers, we performed 876 5 ns simulations and 117 50 ns simulations in total. These simulations take approximately 394,000 CPU hours in total on NERSC Cori Haswell Compute Nodes. This only accounts for around 4.4% of the computation needed to simulate all the polymers from the polymer space with 50 ns simulations, corresponding to a 22.8-fold acceleration. The acceleration would be even larger if we consider that multiple simulations are needed for each polymer to reduce the random errors from MD to match our true prediction accuracy.

Acknowledgments and Disclosure of Funding

This work was supported by Toyota Research Institute. Computational support was provided by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and the Extreme Science and Engineering Discovery Environment, supported by National Science Foundation grant number ACI-1053575. J.L. acknowledges support by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at the Massachusetts Institute of Technology, administered by Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence.

References

[1] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, *et al.*, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," *Nature materials*, vol. 15, no. 10, pp. 1120–1127, 2016.

[2] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, pp. 2224–2232, 2015.

[3] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of computer-aided molecular design*, vol. 30, no. 8, pp. 595–608, 2016.

[4] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," *arXiv preprint arXiv:1704.01212*, 2017.

[5] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature communications*, vol. 8, no. 1, pp. 1–8, 2017.

[6] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," *Physical review letters*, vol. 120, no. 14, p. 145301, 2018.

[7] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet–a deep learning architecture for molecules and materials," *The Journal of Chemical Physics*, vol. 148, no. 24, p. 241722, 2018.

[8] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," *Chemistry of Materials*, vol. 31, no. 9, pp. 3564–3572, 2019.

[9] M. A. Webb, Y. Jung, D. M. Pesko, B. M. Savoie, U. Yamamoto, G. W. Coates, N. P. Balsara, Z.-G. Wang, and T. F. Miller III, "Systematic computational and experimental investigation of lithium-ion transport mechanisms in polyester-based polymer electrolytes," *ACS central science*, vol. 1, no. 4, pp. 198–205, 2015.

[10] B. M. Savoie, M. A. Webb, and T. F. Miller III, "Enhancing cation diffusion and suppressing anion diffusion via lewis-acidic polymer electrolytes," *The journal of physical chemistry letters*, vol. 8, no. 3, pp. 641–646, 2017.

[11] A. France-Lanord, Y. Wang, T. Xie, J. A. Johnson, Y. Shao-Horn, and J. C. Grossman, "Effect of chemical variations in the structure of poly (ethylene oxide)-based polymers on lithium transport in concentrated electrolytes," *Chemistry of Materials*, vol. 32, no. 1, pp. 121–126, 2019.

[12] R. Agrawal and G. Pandey, "Solid polymer electrolytes: materials designing and all-solid-state battery applications: an overview," *Journal of Physics D: Applied Physics*, vol. 41, no. 22, p. 223001, 2008.

[13] K. S. Ngai, S. Ramesh, K. Ramesh, and J. C. Juan, "A review of polymer electrolytes: fundamental, approaches and applications," *Ionics*, vol. 22, no. 8, pp. 1259–1279, 2016.

[14] D. T. Hallinan Jr and N. P. Balsara, "Polymer electrolytes," *Annual review of materials research*, vol. 43, pp. 503–525, 2013.

[15] J. J. Irwin and B. K. Shoichet, "Zinc- a free database of commercially available compounds for virtual screening," *Journal of chemical information and modeling*, vol. 45, no. 1, pp. 177–182, 2005.

[16] S. A. Wildman and G. M. Crippen, "Prediction of physicochemical parameters by atomic contributions," *Journal of chemical information and computer sciences*, vol. 39, no. 5, pp. 868–873, 1999.

[17] "RDKit: Open-source cheminformatics." `http://www.rdkit.org`. [Online; accessed 11-April-2013].

[18] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2017.

[19] B. Du, T. Xinyao, Z. Wang, L. Zhang, and D. Tao, "Robust graph-based semisupervised learning for noisy labeled data via maximum correntropy criterion," *IEEE transactions on cybernetics*, vol. 49, no. 4, pp. 1440–1453, 2018.

A Estimate true prediction error from noisy data

We assume there exists a deterministic function f that maps from the polymer structure $\mathcal{G}$ to its true target property. However, due to the random errors introduced by the initial configuration in MD simulations, the simulated target property t has a small random error ϵ,

$$t = f(\mathcal{G}) + \epsilon, \tag{6}$$

where ϵ follows a normal distribution with zero bias $\mathcal{N}(0, \eta)$. Here, we assume that ϵ is not a function of $\mathcal{G}$, i.e. different polymers have the same random error independent of their structure. This is approximately correct based on the differences in conductivity of the same polymer between two independent MD simulations in the log scale (Fig. 2).

To estimate the true prediction error of our model, we write our graph neural network model as a deterministic function g that predicts polymer property based on their structure $\mathcal{G}$,

$$y = g(\mathcal{G}). \tag{7}$$

Note that we use different labels for the predicted property y and the MD simulated property t.

Under these assumptions, the mean squared error between ML predictions and MD simulated properties is,

$$\mathrm{MSE}(y, t) = \mathop{\mathbb{E}}_{\mathcal{G}}[(y - t)^2] = \mathop{\mathbb{E}}_{\mathcal{G}}[(y - f(\mathcal{G}) - \epsilon)^2] = \mathop{\mathbb{E}}_{\mathcal{G}}[(y - f(\mathcal{G}))^2] + \mathop{\mathbb{E}}_{\mathcal{G}}[\epsilon^2]. \tag{8}$$

Note that in the last step we use the fact that $\mathbb{E}_{\mathcal{G}}[\epsilon] = 0$.

The mean squared error between two independent MD simulations for the same polymer is,

$$\mathrm{MSE}(t_1, t_2) = \mathop{\mathbb{E}}_{\mathcal{G}}[(t_1 - t_2)^2] = \mathop{\mathbb{E}}_{\mathcal{G}}[(\epsilon_1 - \epsilon_2)^2] = 2\mathop{\mathbb{E}}_{\mathcal{G}}[\epsilon^2]. \tag{9}$$

$\mathrm{MSE}(t_1, t_2)$ can be calculated from Fig. 2. Therefore, $\mathbb{E}_{\mathcal{G}}[\epsilon^2] \approx 0.0137$.

The mean squared error between ML predictions and the true target property, i.e. true prediction error, can then be calculated with,

$$\mathrm{MSE}(y, f(\mathcal{G})) = \mathop{\mathbb{E}}_{\mathcal{G}}[(y - f(\mathcal{G}))^2] = \mathrm{MSE}(y, t) - \mathop{\mathbb{E}}_{\mathcal{G}}[\epsilon^2]. \tag{10}$$

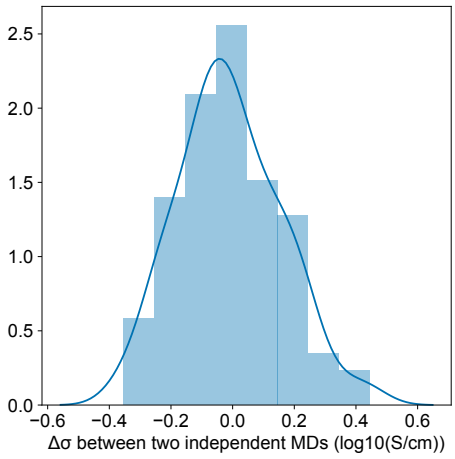

Figure 2: Differences in conductivity of the same polymer between two independent 5 ns molecular dynamics simulations.